Asymptotic Equivalence of Density Estimation and Gaussian White Noise

Michael Nussbaum
Weierstrass Institute, Berlin

September 1995

Abstract

Signal recovery in Gaussian white noise with variance tending to zero has served for some time as a representative model for nonparametric curve estimation, having all the essential traits in a pure form. The equivalence has mostly been stated informally, but an approximation in the sense of Le Cam's deficiency distance Δ would make it precise. The models are then asymptotically equivalent for all purposes of statistical decision with bounded loss. In nonparametrics, a first result of this kind has recently been established for Gaussian regression (Brown and Low, 1993). We consider the analogous problem for the experiment given by n i. i. d. observations having density f on the unit interval. Our basic result concerns the parameter space of densities which are in a Hölder ball with exponent $\alpha > \frac{1}{2}$ and and which are uniformly bounded away from zero. We show that an i. i. d. sample of size n with density f is globally asymptotically equivalent to a white noise experiment with drift $f^{1/2}$ and variance $\frac{1}{4}n^{-1}$. This represents a nonparametric analog of Le Cam's heteroscedastic Gaussian approximation in the finite dimensional case. The proof utilizes empirical process techniques related to the Hungarian construction. White noise models on f and $\log f$ are also considered, allowing for various "automatic" asymptotic risk bounds in the i. i. d. model from white noise. As first applications we discuss exact constants for L_2 and Hellinger loss.

1 Introduction and main result

One of the basic principles of Le Cam's (1986) asymptotic decision theory is to approximate general experiments by simple ones. In particular, *weak convergence to Gaussian shift experiments* has now become a standard tool for establishing asymptotic risk bounds. The risk bounds implied by weak convergence are generally estimates from below, and in most of the literature the efficiency of procedures is more or less shown on an ad hoc basis. However, a systematic approach to the attainment problem is also made possible by Le Cam's theory, based on the notion of *strong convergence of experiments* which means proximity in the sense of the full deficiency distance. But due to the inherent technical difficulties of handling the

---

1990 *Mathematics Subject Classification.* Primary 62 G 07 ; Secondary 62 B 15, 62 G 20
Key words and phrases. Nonparametric experiments, deficiency distance, likelihood process, Hungarian construction, asymptotic minimax risk, curve estimation.

deficiency concept, this possibility is rarely made use of, even in root-n consistent parametric problems.

In nonparametric curve estimation models of the "ill posed" class where there is no root-n consistency, research has focused for a long time on optimal rates of convergence. In these problems, limits of experiments for $n^{-1/2}$-localized parameter are not directly useful for risk bounds. But now a theory of *exact asymptotic risk constants* is also developing in the context of slower rates of convergence. Such an exact risk bound was first discovered by Pinsker (1980) in the problem of *signal recovery in Gaussian white noise*, which is by now recognized as the basic or "typical" nonparametric curve estimation problem. The cognitive value of this model had already been realized by Ibragimov and Khasminski (1977). These risk bounds have been established since then in a variety of other problems, e. g. density, nonparametric regression, spectral density, see Efroimovich and Pinsker (1982), Golubev (1984), Nussbaum (1985), and they have also been substantially extended conceptually (Korostelev (1993), Donoho, Johnstone, Kerkyacharian, Picard (1995)). The theory is now at a stage where the approximation of the various particular curve estimation problems by the white noise model could be made formal. An important step in this direction has been made by Brown and Low (1993) by relating Gaussian regression to the signal recovery problem. These models are essentially the continuous and discrete versions of each other. The aim of this paper is to establish the *formal approximation by the white noise model* for the problem of density estimation from an i. i. d. sample.
To formulate our main result, define a basic parameter space Σ of densities as follows. Let for $\alpha \in (0,1)$ and $M > 0$

$$\Lambda^\alpha(M) = \{f : |f(x) - f(y)| \le M \ |x-y|^\alpha, \ x, y \in [0,1]\}$$

be a Hölder ball of functions with exponent α . Define for $\epsilon > 0$ a set $\mathcal{F}_{\ge\epsilon}$ as the set of densities on $[0,1]$ bounded below by ϵ:

$$\mathcal{F}_{\ge\epsilon} = \left\{ f : \ \int_0^1 f = 1, \ f(x) \ge \epsilon, \ x \in [0,1] \right\}. \tag{1}$$

Define an a priori set, for given $\alpha > \frac{1}{2}$, $M > 0$, $\epsilon > 0$,

$$\Sigma_{\alpha,M,\epsilon} = \Lambda^\alpha(M) \cap \mathcal{F}_{\ge\epsilon}. \tag{2}$$

Let Δ be Le Cam's deficiency pseudodistance between experiments having the same parameter space. For the convenience of the reader a formal definition is given in section 10 below. For two sequences of experiments $\mathbb{E}_n$ and $\mathbb{F}_n$ we shall say that they are *asymptotically equivalent* if $\Delta(\mathbb{E}_n, \mathbb{F}_n) \to 0$ as $n \to \infty$. Let dW denote the standard Gaussian white noise process on the unit interval.

1.1 Theorem. *Let Σ be a set of densities contained in $\Sigma_{\alpha,M,\epsilon}$ for some $\epsilon > 0$, $M > 0$ and $\alpha > \frac{1}{2}$. Then the experiments given by observations*

$$y_i, \ i = 1, \dots, n \qquad \text{i. i. d. with density } f \tag{3}$$

$$dy(t) = f^{1/2}(t)dt + \frac{1}{2} n^{-1/2} dW(t), \ t \in [0,1] \tag{4}$$

with $f \in \Sigma$ are asymptotically equivalent.

This result is closely related to Le Cam's global asymptotic normality for parametric models. Let in the i. i. d. model f be in a parametric family $(f_\vartheta,\ \vartheta \in \Theta)$ where $\Theta \subset \mathbb{R}^k$, which is sufficiently regular and has Fisher information matrix $I(\vartheta)$ at point ϑ. Then the i. i. d. model may be approximated by a heteroscedastic Gaussian experiment

$$y = \vartheta + n^{-1/2} I(\vartheta)^{-1/2} \eta \tag{5}$$

where η is a standard normal vector and $\vartheta \in \Theta$. We see that (4) is a nonparametric analog of (5) when ϑ is identified with $f^{1/2}$. Indeed, consider the identity for the Fisher information matrix in the parametric case

$$\left\| f_{\vartheta'}^{1/2} - f_\vartheta^{1/2} \right\|_2^2 = 4^{-1} \langle\ \vartheta' - \vartheta, I(\vartheta)(\vartheta' - \vartheta)\ \rangle + o\left(\left\| \vartheta' - \vartheta \right\|_2^2 \right).$$

Regarding formally $f^{1/2}$ itself as a parameter, we find the corresponding Fisher information to be 4 times the unit operator. But even for parametric families (4) seems to be an interesting form of a global approximation: if $f_\vartheta^{1/2}$ is taken as parameter then the resulting Gaussian model has a simple form. One recognizes that the heteroscedastic nature of (5) derives only from the "curved" nature of a general parametric family within the space of roots of densities. This observation was in fact made earlier by Le Cam (1985). In his theorem 4.3 there he established the homoscedastic global Gaussian approximation for i. i. d. models in the *finite dimensional case*. We give a paraphrase of that result in a specialized form. A set Θ' in $L_2(0,1)$ is said to be of finite metric dimension if there is a number D such that every subset of Θ' which can be covered by an ϵ-ball can be covered by no more than 2^D balls of radius $\epsilon/2$, where D does not depend on ϵ. A set of densities f has this property in Hellinger metric if the corresponding set of $f^{1/2}$ has it in $L_2(0,1)$.

1.2 Proposition (Le Cam). *Let Σ be a set of densities on $[0,1]$ having finite dimension in Hellinger metric and fulfilling a further regularity condition (see section 12). Then the experiments given by observations (3), (4) with $f \in \Sigma$ are asymptotically equivalent.*

The actual formulation in Le Cam (1985) is more abstract and general giving a global asymptotic normality in the i. i. d. case for arbitrary random variables, in particular without assumed existence of densities; but finite dimensionality is essential. This result in its conceptual clarity and potential impact seems not to have been well appreciated by researchers; the heteroscedastic form (5) under classical regularity conditions is somewhat better known (cp. Mammen (1986)).

Our main result can thus be viewed as an extension of Le Cam's proposition 1.2 to a nonparametric setting. The value 1/2 of the Hölder exponent α is a critical one, according to a recent result of Brown and Zhang (1995).

White noise models with *fixed* variance do occur as local limits of experiments in root-n consistent nonparametric problems (Millar (1979)), and, via specific renormalizations, also in non root-n consistent curve estimation (Low (1992), Donoho and Low (1992)). Thus various central limit theorems for i. i. d. experiments can be embedded in a relatively simple and closed form approximation by (4). Moreover, for the density f itself and for $\log f$ we also give Gaussian approximations which are "heteroscedastic" in analogy to (5), see remark 2.9, corollary 3.3 below.

The paper is organized as follows. The basic results are developed in an overview fashion in sections 2-4 which may suffice for a first reading. By default, proofs or technical comments for all statements are to be found in part II, i. e. the proof sections 5-12.

In section 2 we develop the basic approximation of likelihood ratios over shrinking neighborhoods of a given density f_0. These neighborhoods $\Sigma_n(f_0)$ are already "nonparametric", in the sense of shrinking slower than $n^{-1/2}$. For proving this, we partition the sample space $[0,1]$ into small intervals and obtain a product experiment structure via poissonization. The Gaussian approximation is then argued via the "space local" empirical process on the small intervals; piecing this together on $[0,1]$ yields the basic parameter-local Gaussian approximation over $f \in \Sigma_n(f_0)$. Once in a Gaussian framework, we manipulate likelihood ratios to obtain other approximations, in particular the one with trend $f^{1/2}$. For these experiments which are all Gaussian we use the methodology of Brown and Low (1993), who did compare the white noise model with its discrete version (the Gaussian regression model).
It remains to piece together the parameter-local approximations by means of a preliminary estimator; this is the subject of section 3. Our method of globalization is somewhat different from Le Cam's which works in the parametric case; the concept of metric entropy or dimension and related theory are not utilized. But obviously these methods which already proved fruitful in nonparametrics (Birgé (1983), Van de Geer (1990)) have a potential application also here. Some statistical consequences are discussed in section 4; here we focus on exact constants for L_2-loss. As an exercise we derive the result of Efroimovich and Pinsker (1982) on density estimation from the white noise model; simultaneously we extend it and give a variant for Hellinger loss.
As a basic text for the asymptotic theory of experiments we refer to Strasser (1985). We use C as a generic notation for positive constants; for sequences the symbol $a_n \asymp b_n$ means the usual equivalence in rate, while $a_n \sim b_n$ means $a_n = b_n(1+o(1))$.

2 The local approximation

Our first Gaussian approximation will be established in a parameter local framework. Suppose we have i. i. d. observations y_i, $i = 1, \dots, n$ with distribution P_f having Lebesgue density f on the interval $[0,1]$, and it is known a priori that f belongs to a set of densities Σ. Henceforth in the paper we will set $\Sigma = \Sigma_{\alpha,M,\epsilon}$ for some $\epsilon > 0$, $M > 0$ and $\alpha > 1/2$.
Let $\| \cdot \|_p$ denote the norm in the space $L_p(0,1)$, $1 \le p \le \infty$. Let γ_n be the sequence

$$\gamma_n = n^{-1/4}(\log n)^{-1}, \tag{6}$$

and for any $f_0 \in \Sigma$ define a class $\Sigma_n(f_0)$ by

$$\Sigma_n(f_0) = \left\{ f \in \Sigma : \ \left\| \frac{f}{f_0} - 1 \right\|_\infty \le \gamma_n \right\}. \tag{7}$$

For given $f_0 \in \Sigma$ we define a local (around f_0) product experiment

$$\mathbb{E}_{0,n}(f_0) = \left([0,1]^n, \mathcal{B}^n_{[0,1]}, (P_f^{\otimes n}, \ f \in \Sigma_n(f_0)) \right). \tag{8}$$

Let F_0 be the distribution function corresponding to f_0 and let

$$K(f_0 \parallel f) = -\int \log \frac{f}{f_0} dF_0$$

be the Kullback-Leibler relative entropy. Let W be the standard Wiener process on $[0,1]$ and consider an observed process

$$y(t) = \int_0^t \log \frac{f}{f_0}(F_0^{-1}(u))du + tK(f_0 \| f) + n^{-1/2}W(t), \ t \in [0,1]. \tag{9}$$

Let Q_{n,f,f_0} be the distribution of this process on the function space $C_{[0,1]}$ equipped with its Borel σ-algebra $\mathcal{B}_{C_{[0,1]}}$, and

$$\mathbb{E}_{1,n}(f_0) = \Big(C_{[0,1]}, \mathcal{B}_{C_{[0,1]}}, (Q_{n,f,f_0},\ f\ \in\ \Sigma_n(f_0))\Big) \tag{10}$$

be the corresponding experiment when f varies in the neighborhood $\Sigma_n(f_0)$.

2.1 Theorem. *Define $\Sigma_n(f_0)$ as in (7), (6). Then*

$$\Delta(\mathbb{E}_{0,n}(f_0), \mathbb{E}_{1,n}(f_0)) \longrightarrow 0 \text{ as } \ n\ \longrightarrow \infty$$

uniformly over $f_0\ \in\ \Sigma$.

The proof is based upon the following principle, described in Le Cam and Yang (1991), p. 16. Consider two experiments $\mathbb{E}_i = (\Omega_i, \mathcal{A}_i, (P_{i,\vartheta}, \vartheta \in \Theta))$, $i = 0, 1$ having the same parameter set Θ. Assume there is some point $\vartheta_0 \in \Theta$ such that all the $P_{i,\vartheta}$ are dominated by P_{i,ϑ_0}, $i = 0, 1$ and form $\Lambda_i(\vartheta) = dP_{i,\vartheta}/dP_{i,\vartheta_0}$. Consider $\Lambda_i = (\Lambda_i(\vartheta), \vartheta \in \Theta)$ as stochastic processes indexed by ϑ given on the probability space $(\Omega_i, \mathcal{A}_i, P_{i,\vartheta_0})$. By a slight abuse of language, we call these the *likelihood processes* of the experiments $\mathbb{E}_i$ (note that the distribution is taken under P_{i,ϑ_0} here). Suppose also that there are versions Λ_i^* of these likelihood processes defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$.

2.2 Proposition. *The deficiency distance $\Delta(\mathbb{E}_1, \mathbb{E}_2)$ satisfies*

$$\Delta(\mathbb{E}_0, \mathbb{E}_1) \leq \sup_{\vartheta \in \Theta} E_{\mathbb{P}} |\Lambda_0^*(\vartheta) - \Lambda_1^*(\vartheta)|.$$

Proof. It is one of the basic facts of Le Cam's theory that for dominated experiments, the equivalence class is determined by the distribution of the likelihood processes under P_{i,ϑ_0} when ϑ_0 is assumed fixed. This means that in the above framework, we have $\Delta(\mathbb{E}_0, \mathbb{E}_1) = 0$ iff $\mathcal{L}(\Lambda_0 | P_{0,\vartheta_0}) = \mathcal{L}(\Lambda_1 | P_{1,\vartheta_0})$. Thus, if we construct an experiment $\mathbb{E}_i^*$ with likelihood process Λ_i^*, we obtain equivalence: $\Delta(\mathbb{E}_i, \mathbb{E}_i^*) = 0$. The random variables $\Lambda_i^*(\vartheta)$ on $(\Omega, \mathcal{A}, \mathbb{P})$ have the same distributions as $\Lambda_i(\vartheta)$ on $(\Omega_i, \mathcal{A}_i, P_{i,\vartheta_0})$, for all $\vartheta \in \Theta$; hence they are positive and integrate to one. They may hence be considered as $\mathbb{P}$-densities on $(\Omega, \mathcal{A})$, indexed by ϑ. These densities define measures $P_{i,\vartheta}^*$ on $(\Omega, \mathcal{A})$, and experiments $\mathbb{E}_i^* = (\Omega, \mathcal{A}, (P_{i,\vartheta}^*,\ \vartheta \in \Theta))$, $i = 0, 1$. By construction, the likelihood process for $\mathbb{E}_i^*$ is $\Lambda_i^*(\vartheta)$, so $\Delta(\mathbb{E}_i, \mathbb{E}_i^*) = 0$, $i = 0, 1$. Hence $\Delta(\mathbb{E}_0, \mathbb{E}_1) = \Delta(\mathbb{E}_0^*, \mathbb{E}_1^*)$, and $\mathbb{E}_0^*, \mathbb{E}_1^*$ are given *on the same measurable space* $(\Omega, \mathcal{A})$. In this case, an upper bound for the deficiency distance is

$$\Delta(\mathbb{E}_0^*, \mathbb{E}_1^*) \leq \sup_{\vartheta \in \Theta} \left\| P_{0,\vartheta}^* - P_{1,\vartheta}^* \right\|$$

where $\|\cdot\|$ is the total variation distance between measures (in section 10, (72) take the identity map as a transition M). But $\left\| P_{0,\vartheta}^* - P_{1,\vartheta}^* \right\|$ coincides with $E_{\mathbb{P}} |\Lambda_0^*(\vartheta) - \Lambda_1^*(\vartheta)|$ which is just a L_1-distance between densities. $\square$

The argument may be summarized as follows: versions Λ_i^* of the likelihood processes on a common probability space generate (equivalent) versions of the experiments on a common measurable space for which $\Lambda_i^*(\vartheta)$ are densities. Their L_1-distance bounds the deficiency.

When $\Lambda_i^*(\vartheta)$ are considered as densities it is natural to employ also their Hellinger distance $H(\cdot,\cdot)$; extending notation we will write

$$H^2(\Lambda_0^*(\vartheta),\Lambda_1^*(\vartheta)) = E_{\mathbb{P}}\left((\Lambda_0^*(\vartheta))^{1/2} - (\Lambda_1^*(\vartheta))^{1/2}\right)^2. \tag{11}$$

Making use of the general relation of Hellinger to L_1-distance we obtain

$$\Delta^2(\mathbb{E}_0^*,\mathbb{E}_1^*) \le \sup_{\vartheta\in\Theta} \quad H^2(\Lambda_0^*(\vartheta),\Lambda_1^*(\vartheta)). \tag{12}$$

In the sequel we will work basically with this relation to establish asymptotic equivalence. For our problem, we identify $\vartheta = f, \vartheta_0 = f_0$, $\Theta = \Sigma_n(f_0)$, $P_{0,\vartheta} = P_f^{\otimes n}$, $P_{1,\vartheta} = Q_{n,f,f_0}$. Furthermore, we represent the observations y_i as $y_i = F^{-1}(z_i)$, where z_i are i. i. d. uniform (0,1) random variables and F is the distribution function for the density f (note that F is strictly monotone for $f \in \Sigma$). Let U_n be the empirical process of $z_1,\ldots,z_n$, i. e.

$$U_n(t) = \frac{1}{\sqrt{n}}\sum_{i=1}^n(\chi_{[0,t]}(z_i) - t),\ t\in[0,1].$$

Note that $\mathbb{E}_{0,n}(f_0)$ is dominated by $P_{f_0}^{\otimes n}$; then the likelihood process is

$$\Lambda_{0,n}(f,f_0) = \exp\sum_{i=1}^n \log\left\{\frac{f}{f_0}(F_0^{-1}(z_i))\right\}.$$

Defining

$$\lambda_{f,f_0}(t) = \log\left\{\frac{f}{f_0}(F_0^{-1}(t))\right\} \tag{13}$$

and observing that

$$\int \lambda_{f,f_0}(t)dt = -K(f_0\|f)$$

we then have the following representation:

$$\Lambda_{0,n}(f,f_0) = \exp\left\{n\int\lambda_{f,f_0}(t)\frac{1}{\sqrt{n}}\,U_n(dt) - nK(f_0\|f)\right\}. \tag{14}$$

This suggests a corresponding Gaussian likelihood process: substitute U_n by a Brownian bridge B and renormalize to obtain integral one. We thus form for a uniform $(0,1)$ random variable Z

$$\Lambda_{1,n}(f,f_0) = \exp\left\{n\int\lambda_{f,f_0}(t)\frac{1}{\sqrt{n}}\,B(dt) - \frac{n}{2}\mathrm{Var}\,(\lambda_{f,f_0}(Z))\right\}. \tag{15}$$

For an appropriate standard Wiener process W we have

$$\int\lambda_{f,f_0}(t)\,B(dt) = \int(\lambda_{f,f_0}(t) + K(f_0\,\|\,f))\,W(dt).$$

By rewriting the likelihood process $\Lambda_{1,n}(f,f_0)$ accordingly we see that it corresponds to observations (9) or equivalently to

$$dy(t) = (\lambda_{f,f_0}(t) + K(f_0\|f))dt + n^{-1/2}dW(t),\ t\in[0,1], \tag{16}$$

at least when the parameter space is Σ. Thus $\Lambda_{1,n}(f, f_0)$ is in fact the likelihood process for $\mathbb{E}_{1,n}(f_0)$ in (10).

To find nearby versions of these likelihood processes, fulfilling

$$\sup_{f \in \Sigma_n(f_0)} H^2\left(\Lambda^*_{0,n}(f, f_0), \Lambda^*_{1,n}(f, f_0)\right) \to 0 \tag{17}$$

it would be natural to look for versions of U_n and B on a common probability space ($\mathbb{U}_n$ and $\mathbb{B}_n$, say) which are close, such as in the classical *Hungarian construction* (see Shorack, Wellner (1986), chap. 12, section 1, theor. 2). However the classical Hungarian construction (Komlos-Major-Tusnady inequality) gives an estimate of the uniform distance $\|\mathbb{U}_n - \mathbb{B}_n\|_\infty$ which for our purpose is not optimal. The reason is that the uniform distance may be construed as

$$\|\mathbb{U}_n - \mathbb{B}_n\|_\infty = \sup_{g \in \mathcal{G}} |\mathbb{U}_n(g) - \mathbb{B}_n(g)|$$

where $\mathcal{G}$ is a class of indicators of subintervals of $[0, 1]$. Considering more general classes of functions $\mathcal{G}$ leads to *functional* KMT type results (see Koltchinskii (1994), Rio (1994)). But for an estimate (17) we need to control the random difference $\mathbb{U}_n(g) - \mathbb{B}_n(g)$ only for one given function (λ_{f,f_0} in this case), with a supremum over a function class only after taking expectations (cp the remark on p. 16 of Le Cam and Yang (1991)). Thus for our purpose we ought to use a functional KMT type inequality for a one element function class $\mathcal{G} = \{g\}$, but where the same constants and *one Brownian bridge* are still available over a class of smooth g. Such a result is provided by Koltchinskii (1994), theorem 3.5. We present a version slightly adapted for our purpose. Let $\mathcal{L}_2[0, 1]$ be the space of all square integrable measurable functions on $[0, 1]$ and let $\|\cdot\|_{H_2^{1/2}}$ be the seminorm associated with a Hölder condition with exponent $1/2$ in the L_2-sense (see section 6 for details).

2.3 Proposition. *There are a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and a number C such that for all n, there are versions of the uniform empirical process $\mathbb{U}_n(g)$ and of the Brownian bridge $\mathbb{B}_n(g)$, $g \in \mathcal{L}_2[0, 1]$ such that for all g with $\|g\|_\infty < \infty$, $\|g\|_{H_2^{1/2}} < \infty$ and for all $t \geq 0$*

$$\mathbb{P}(n^{1/2} |\mathbb{U}_n(g) - \mathbb{B}_n(g)| \geq C\, (\|g\|_\infty + \|g\|_{H_2^{1/2}})(t + \log n) \log^{1/2} n) \leq C\, \exp(-t).$$

Specializing $g = \lambda_{f,f_0} - \int \lambda_{f,f_0}$ we come close to establishing the relation (17) for the likelihood processes, but we need an assumption that the neighborhoods $\Sigma_n(f_0)$ shrink with rate $o(n^{-1/3})$. Comparing with the usual nonparametric rates of convergence, we see that such a result is useful only for smoothness $\alpha > 1$. To treat the case $\alpha > 1/2$ however we need neighborhoods of size $o(n^{-1/4})$.
To obtain such a result, it is convenient, rather than using the Hungarian construction globally on $[0, 1]$, to subdivide the interval and use a corresponding independence structure (approximate or exact) of both experiments. In this connection the following result is useful (see Strasser (1985), lemma 2.19).

2.4 Lemma. *Suppose that P_i, Q_i are probability measures on a measurable space $(\Omega_i, \mathcal{A}_i)$, for $i = 1, \ldots, k$. Then*

$$H^2(\bigotimes_{i=1}^k P_i, \bigotimes_{i=1}^k Q_i) \leq 2 \sum_{i=1}^k H^2(P_i, Q_i).$$

Consider a partition of $[0,1]$ into subintervals D_j. The Gaussian experiment $\mathbb{E}_{1,n}(f_0)$ has a convenient independence structure: in the representation (16), observations on the signal $\lambda_{f,f_0}(t) + K(f_0\|f)$ are independent on different pieces D_j. A corresponding approximate product structure for the iid experiment $\mathbb{E}_{0,n}(f_0)$ will be established by Poissonization. Let $\mathbb{E}_{0,j,n}(f_0)$ be the experiment given by observing "interval censored" observations

$$y_i \chi_{D_j}(y_i),\ y_i \text{ i. i. d. with density } f,\ i = 1, \dots, n \tag{18}$$

with $f \in \Sigma_n(f_0)$. We use the symbol $\bigotimes$ for products of experiments having the same parameter space.

2.5 Proposition. *Let k_n be a sequence with $k_n \to \infty$, and consider a partition $D_j = [(j-1)/k_n, j/k_n)$, $j = 1, \dots, k_n$. Then*

$$\Delta(\mathbb{E}_{0,n}(f_0), \bigotimes_{j=1}^{k_n} \mathbb{E}_{0,j,n}(f_0)) \to 0$$

uniformly over $f_0 \in \Sigma$.

Our choice of k_n will be

$$k_n \sim n^{1/2}/(\log n)^2. \tag{19}$$

For each D_j we form a local likelihood process $\Lambda_{0,j,n}(f, f_0)$, as the likelihood process for observations in (18) for given j, and establish a Gaussian approximation like (17) with a rate. Let $A_j = F_0(D_j)$ and let $\mathbb{E}_{1,j,n}(f_0)$ be the Gaussian experiment

$$dy(t) = \chi_{A_j}(t)\ (\lambda_{f,f_0}(t) + K(f_0\|f))dt + n^{-1/2} dW(t),\ t \in [0,1] \tag{20}$$

with parameter space $\Sigma_n(f_0)$. Let $\Lambda_{1,j,n}(f, f_0)$ be the corresponding likelihood process.

2.6 Proposition. *On the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ of proposition 2.3, there are versions $\Lambda^*_{i,j,n}(f, f_0)$, $i = 0, 1$ such that*

$$\sup_{f \in \Sigma_n(f_0)} H^2\left(\Lambda^*_{0,j,n}(f, f_0), \Lambda^*_{1,j,n}(f, f_0)\right) = O(\gamma_n^2 (\log n)^3) \tag{21}$$

uniformly over $j = 1, \dots, k_n$ and $f_0 \in \Sigma$.

This admits the following interpretation. Define $m_n = n/k_n$; in our setting this is the stochastic order of magnitude of the number of observations y_i falling into D_j. Thus for the local likelihood process $\Lambda_{0,j,n}(f, f_0)$ the number m_n represents an "effective sample size" in a rate sense. In view of (6) and (19) we have $\gamma_n \sim m_n^{-1/2}$, and since this is the shrinking rate of $\Sigma_n(f_0)$ in the uniform norm, it is also the shrinking rate of this set of densities restricted to D_j, and of the corresponding set of conditional densities. Thus in a sense we are in a classical setting with sample size m_n and a root-m_n shrinking neighborhood. The result (21) implies

$$\Delta(\mathbb{E}_{0,j,n}(f_0), \mathbb{E}_{1,j,n}(f_0)) = O(m_n^{-1/2} (\log n)^{3/2}), \tag{22}$$

i. e. we have a root-m_n rate up to a log-term. Note that here we have introduced a "space local" aspect in addition to the already present parameter local one. In piecing together these space local approximations, we will crucially use the product measure estimate of lemma 2.4.

This motivates our choice to work with the Hellinger distance, for the likelihood processes construed as densities.

Proof of theorem 2.1. The Gaussian experiment $\mathbb{E}_{1,n}(f_0)$ decomposes exactly:

$$\Delta(\mathbb{E}_{1,n}(f_0), \bigotimes_{j=1}^{k_n} \mathbb{E}_{1,j,n}(f_0)) = 0.$$

According to (12) and lemma 2.4 we have

$$\Delta^2(\bigotimes_{j=1}^{k_n} \mathbb{E}_{0,j,n}(f_0), \bigotimes_{j=1}^{k_n} \mathbb{E}_{1,j,n}(f_0)) \leq 2 \sup_{f \in \Sigma_n(f_0)} \sum_{j=1}^{k_n} H^2\left(\Lambda^*_{0,j,n}(f, f_0), \Lambda^*_{1,j,n}(f, f_0)\right)$$

By proposition 2.6 this is bounded by

$$O\left(k_n \gamma_n^2 (\log n)^3\right) = O\left((\log n)^{-1}\right) = o(1),$$

and these estimates hold uniformly over $f_0 \in \Sigma$. $\square$

Low (1992) considered experiments given by local (on D_j) perturbations of a fixed density f_0 and applied a local asymptotic normality argument to obtain strong convergence to a Gaussian experiment. This amounts to having (22) without a rate, and it is already useful for a number of nonparametric decision problems, like estimating the density at a point. Golubev (1991) used a similar argument for treating estimation in L_2-loss.
We are now able to identify several more asymptotically equivalent models. This is based on the following reasoning, applied by Brown and Low (1993) to compare Gaussian white noise models. Consider the measure of the process $n^{-1/2}\ W(t)$, $t \in [0,1]$ shifted by a function $\int_0^t g_i$, $i = 1, 2$, where $g_i \in \mathcal{L}_2[0,1]$; call these measures P_i. Then

$$H^2(P_1, P_2) = 2\left(1 - \exp\left\{-\frac{n}{8}\ \|g_1 - g_2\|_2^2\right\}\right). \tag{23}$$

If $(g_{i,\vartheta},\ \vartheta \in \Theta)$, $i = 1, 2$ are two parametric families then the respective experiments are asymptotically equivalent if $\|g_{1,\vartheta} - g_{2,\vartheta}\|_2 = o(n^{-1/2})$ uniformly over $\vartheta \in \Theta$. In the Gaussian experiment $\mathbb{E}_{1,n}(f_0)$ of (16), the shift is essentially a log-density ratio. We know that $\log(f/f_0)$ is small over $f \in \Sigma_n(f_0)$; expanding the logarithm we get asymptotically equivalent experiments with parameter space $\Sigma_n(f_0)$.
Accordingly, let $\mathbb{E}_{2,n}(f_0)$ be the experiment given by observations

$$dy(t) = (f(t) - f_0(t))dt + n^{-1/2} f_0^{1/2}(t) dW(t),\ t \in [0,1] \tag{24}$$

with parameter space $\Sigma_n(f_0)$, and let $\mathbb{E}_{3,n}(f_0)$ correspondingly be given by

$$dy(t) = (f^{1/2}(t) - f_0^{1/2}(t))dt + \frac{1}{2} n^{-1/2} dW(t),\ t \in [0,1]. \tag{25}$$

2.8 Theorem. *The experiments $\mathbb{E}_{i,n}(f_0)$, $i = 1, 2, 3$ are asymptotically equivalent, uniformly over $f_0 \in \Sigma$.*

2.9 Remark. The equivalence class of $\mathbb{E}_{1,n}(f_0)$ is not changed when the additive term $-f_0(t)dt$ in (24) is omitted, since this term does not depend on the parameter f, and omitting it amounts to a translation of the observed process y by a known quantity. Moreover, in the proof below it will be seen that in the representation (16) of $\mathbb{E}_{1,n}(f_0)$ the term $K(f_0\|f)dt$ is asymptotically negligible. Analogous statements are true for the other variants; hence locally asymptotically equivalent experiments for $f \in \Sigma_n(f_0)$ (with uniformity over $f_0 \in \Sigma$) are also given by

$$y_i,\ i = 1, \dots, n \qquad \text{i. i. d. with density } f \tag{26}$$

$$dy(t) = \log f(F_0^{-1}(t))dt + n^{-1/2}dW(t),\ t \in [0,1] \tag{27}$$

$$dy(t) = f(t)dt + n^{-1/2} f_0^{1/2}(t)dW(t),\ t \in [0,1] \tag{28}$$

$$dy(t) = f^{1/2}(t)dt + \frac{1}{2}n^{-1/2}dW(t),\ t \in [0,1]. \tag{29}$$

□

Note that (28) is related to the weak convergence of the empirical distribution function $\bar{F}_n$

$$n^{1/2}\left(\bar{F}_n - F\right) \Rightarrow B \circ F.$$

Indeed, arguing heuristically, when F is in a shrinking neighborhood of F_0 we have $B \circ F \approx B \circ F_0$, while $\bar{F}_n$ is a sufficient statistic. We obtain

$$\bar{F}_n \approx F + n^{-1/2} B \circ F_0$$

which suggests a Gaussian accompanying experiment (28). This reasoning is familiar as a heuristic introduction to limiting Gaussian shift experiments, when neighborhoods are shrinking with rate $n^{-1/2}$. However our neighborhoods $f \in \Sigma_n(f_0)$ are larger (recall $\gamma_n = n^{-1/4}(\log n)^{-1}$).

3 From local to global results

The local result concerning a shrinking neighborhood of some f_0 is of limited value for statistical inference since in general such prior information cannot be assumed. Following Le Cam's general principles, we shall construct an experiment where the prior information is furnished by a preliminary estimator, and subsequently the local Gaussian approximation is built around the estimated parameter value.
To formalize this approach, let N_n define a "fraction of the sample size", i. e. N_n is a sequence $N_n \to \infty$, $N_n < n$, and consider the corresponding fraction of the sample $y_1, \dots, y_{N_n}$. Let then $\hat{f}_n$ be an estimator of f based on this fraction, fulfilling (with $P_{n,f}$ the pertaining measure)

$$\inf_{f \in \Sigma} P_{n,f}(\hat{f}_n \in \Sigma_n(f)) \longrightarrow 1. \tag{30}$$

The set Σ must be such that the shrinking rate of $\Sigma_n(f)$ is an attainable rate for estimators. If f has a bounded derivative of order α, we have for f an attainable rate in sup-norm $(n/\log n)^{-\alpha/(2\alpha+1)}$ (see Woodrofe (1967)). The required sup norm rate is $\gamma_n = o(n^{-1/4})$; this corresponds to $\alpha > 1/2$. Thus we may expect for the Hölder smoothness classes assumed

here that the rate γ_n is attainable if the size N_n of the fraction is sufficiently large. We will allow for a range of choices:

(31) $$n/\log n \leq N_n \leq n/2.$$

Define $\mathbb{E}_{0,n}$ to be the original i. i. d. experiment (3) with global parameter space Σ.

3.1 Lemma. *Suppose (31) holds. Then in $\mathbb{E}_{0,n}$ there exists a sequence of estimators $\hat{f}_n$ depending only on $y_1, \ldots, y_{N_n}$ fulfilling (30). One may assume that for each n, the estimator takes values in a finite set of functions in Σ.*

The following construction of a global approximating experiment assumes such an estimator sequence fixed. The idea is to substitute $\hat{f}_n$ for f_0 in the local Gaussian approximation and to retain the first fraction of the i. i. d. sample. Recall that our local Gaussian approximations were given by families $(Q_{n,f,f_0},\ f \in \Sigma_n(f_0))$, cp. (10). Note that $f \in \Sigma_n(f_0)$ is essentially the same as $f_0 \in \Sigma_n(f)$. Accordingly we now consider the event $\hat{f}_n \in \Sigma_n(f)$, and let f range in the unrestricted parameter space Σ. We look at the second sample part, of size $n - N_n$, with its initial i. i. d. family $(P_f^{\otimes(n-N_n)},\ f \in \Sigma)$. Based on the results of the previous section, we can hope that this family will be close, in the experiment sense, to the conditionally Gaussian family $(Q_{n-N_n,f,\hat{f}_n},\ f \in \Sigma)$, on the event $\hat{f}_n \in \Sigma_n(f)$. The measures $Q_{n,f,\hat{f}_n}$, which now depend on $\hat{f}_n$, have to be interpreted as conditional measures, and we form a joint distribution with the first sample fraction.
This idea is especially appealing when the locally approximating Gaussian measure Q_{n,f,f_0} does not depend on the "center" f_0. In this case the resulting global experiment will have a convenient product structure, as we shall see. This is the case with the variant (29) in remark 2.9, when we parametrize with $f^{1/2}$.
To be more precise, define Q_{i,n,f,f_0}, $i = 1, 2, 3$ to be the distributions of $(y(t),\ t \in [0,1])$ in (27), (28), (29). Consider a "compound experiment" given by joint observations $y_1, \ldots, y_{N_n}$ and $y = (y(t),\ t \in [0,1])$, where

(32) $$y_1, \ldots, y_{N_n} \text{ i. i. d. with density } f$$

(33) $$\mathcal{L}(y | y_1, \ldots, y_{N_n}) = Q_{i,n-N_n,f,\hat{f}_n}.$$

Here (33) describes the conditional distribution of y given $y_1, \ldots, y_{N_n}$. Define $R_{i,n,f}(\hat{f})$ to be the joint distribution of $y_1, \ldots, y_{N_n}$ and y in this setup, for $i = 1, 2, 3$; the notation signifies dependence on the sequence of decision functions $\hat{f} = \{\hat{f}_n\}_{n \geq 1}$ (not dependence on the estimator value). Then the compound experiment is

$$\mathbb{E}_{i,n}(\hat{f}) = \left([0,1]^n \times C_{[0,1]}, \mathcal{B}^n_{[0,1]} \otimes \mathcal{B}_{C_{[0,1]}}, (R_{i,n,f}(\hat{f}),\ f \in \Sigma)\right).$$

Since $Q_{3,n,f,f_0} = Q_{3,n,f}$ does not depend on f_0, the measure $R_{3,n,f}(\hat{f}) = R_{3,n,f}$ does not depend on $\hat{f}$ either, and is just the product measure of $P_f^{\otimes n} \otimes Q_{3,n-N_n,f}$. We also write $\mathbb{E}_{3,n}(\hat{f}) = \mathbb{E}_{3,n}$. The technical implementation of the above heuristic reasoning (see section 10) gives the following result.

3.2 Theorem. *Suppose (31) holds and let $\hat{f}_n$ be a sequence of estimators as in lemma 3.1. Then for $i = 1, 2, 3$,*

$$\Delta(\mathbb{E}_{0,n}, \mathbb{E}_{i,n}(\hat{f})) \longrightarrow 0.$$

To restate this in a more transparent fashion, we refer to $y_1, \ldots, y_{N_n}$ and $y = (y(t),\ t \in [0,1])$ in (32), (33) as the first and second parts of the compound experiment, respectively. Let $\hat{F}_n$ be the distribution function corresponding to the realized density estimator $\hat{f}_n$.

3.3 Corollary. *Under the conditions of theorem 3.3, the compound experiments with first part*

(34) $$y_i,\ i = 1, \ldots, N_n \qquad \textit{i. i. d. with density } f$$

and respective second parts

(35) $$y_i,\ i = N_n + 1, \ldots, n \qquad \textit{i. i. d. with density } f$$

(36) $$dy(t) = \log f(\hat{F}_n^{-1}(t)) + (n - N_n)^{-1/2} dW(t),\ t \in [0,1]$$

(37) $$dy(t) = f(t)dt + (n - N_n)^{-1/2} \hat{f}_n^{1/2}(t) dW(t),\ t \in [0,1]$$

(38) $$dy(t) = f^{1/2}(t)dt + \frac{1}{2}(n - N_n)^{-1/2} dW(t),\ t \in [0,1]$$

with $f \in \Sigma$ are all asymptotically equivalent.

For obtaining a closed form global approximation, the compound experiment $\mathbb{E}_{3,n}$, i. e. (34), (38), is the most interesting one, in view of its product structure and independence of $\hat{f}$. Here the estimator sequence $\hat{f}$ only serves to show asymptotic equivalence to $\mathbb{E}_{0,n}$; it does not show up in the target experiment $\mathbb{E}_{3,n}$ itself. This structure of $\mathbb{E}_{3,n}$ suggests to employ an estimator based on the second part to move on.

3.4 Lemma. *Suppose (31) holds. Then in $\mathbb{E}_{3,n}$ there exists a sequence of estimators $\check{f}_n$ depending only on y in (38) fulfilling (30). The second statement of lemma 3.1 also applies.*

Note the symmetry to lemma 3.1. Here we exploit the well known parallelism of density estimation and white noise on the rate of convergence level.

Proof of theorem 1.1. We choose $N_n = [n/2]$. On the resulting compound experiment $\mathbb{E}_{3,n}$ we may then operate again, reversing the roles of first and second part. We may in turn substitute $y_1, \ldots, y_{N_n}$ by a white noise model, using a preliminary estimator based on (38). The existence of such an estimator is guaranteed by the previous lemma. Thus substituting $y_1, \ldots, y_{N_n}$ by white noise leads to an experiment with joint observations

$$\begin{aligned} dy_1(t) &= f^{1/2}(t)dt + \frac{1}{2} N_n^{-1/2} dW_1(t),\ t \in [0,1] \\ dy_2(t) &= f^{1/2}(t)dt + \frac{1}{2}(n - N_n)^{-1/2} dW_2(t),\ t \in [0,1]. \end{aligned}$$

where W_1, W_2 are independent Wiener processes. A sufficiency argument shows this equivalent to observing n i. i. d. processes, each distributed as

$$dy(t) = f^{1/2}(t)dt + \frac{1}{2} dW(t),\ t \in [0,1],$$

which in turn is equivalent to (4). □

4 An application: exact constants for L_2-risk

Let $\mathcal{F}_n \subset \Sigma$ be any a priori set for the density f, and l_n be a bounded loss function in an estimation problem:

$$l_n(g,f) \le C \quad \text{for} \quad f \in \mathcal{F}_n \qquad \text{and for all possible estimator values g.}$$

Let as before $\mathbb{E}_{0,n}$ be the density experiment with full parameter space Σ, and let $\rho_{0,n}(l_n, \mathcal{F}_n)$ be the minimax risk there for restricted parameter space $\mathcal{F}_n$ and loss function l_n. Let $\mathbb{E}_{\sim,n}$ be another experiment with parameter space Σ, and let $\rho_{\sim,n}(l_n, \mathcal{F}_n)$ be the analogous minimax risk there.

4.1 Proposition. *Let l_n be a uniformly bounded sequence of loss functions. Suppose that $\Delta(\mathbb{E}_{0,n}, \mathbb{E}_{\sim,n}) \to 0$. Then for any sequence of parameter spaces $\mathcal{F}_n \subset \Sigma$ the minimax risks fulfill*

$$\rho_{0,n}(l_n, \mathcal{F}_n) - \rho_{\sim,n}(l_n, \mathcal{F}_n) \to 0.$$

In particular one may consider loss functions l_n such as

$$l_n(g,f) = l\left(n^{1-r} \|g - f\|_2^2\right) \tag{39}$$

where n^{r-1} is the optimal rate of convergence for squared L_2-loss and l is a bounded function. Let $\mathfrak{L}$ be the class of continuous nondecreasing functions on $[0,\infty)$ such that $0 \le l(x) \le x$, $x \in [0,\infty)$, and let $\mathfrak{L}_b$ be the class of bounded $l \in \mathfrak{L}$.
The exact risk asymptotics over Sobolev classes for squared L_2-risk (i.e. for an unbounded $l(x) = x$) was found by Pinsker (1980) for white noise; it was subsequently carried over to density estimation by Efroimovich and Pinsker (1982). Tsybakov (1994) generalized the Pinsker result to bounded l; this is particularly suitable for an argument via equivalence. As an exercise let us deduce the density case result for bounded loss from the white noise approximation.
We begin by stating Pinsker's minimax risk bound in a very simple Gaussian model, which is instructive for understanding the general case. Consider observations

$$y_j = f_{(j)} + \xi_j, \ j = 1, \dots, n \tag{40}$$

where ξ_j are independent standard normal, and the vector $f = (f_{(j)})_{j=1,\dots,n}$ is in a set

$$W_n = \left\{ f \in \mathbb{R}^n : n^{-1} \|f\|^2 \le 1 \right\}.$$

where $\|\cdot\|$ is euclidean norm. Denote this experiment by $\mathbb{E}^0_{\sim,n}$. Consider a loss function

$$l_n(g,f) = l\left(n^{-1} \|g - f\|^2\right) \tag{41}$$

and let $\rho_{\sim,n}(l_n, W_n)$ be the minimax risk over all estimators, for parameter space W_n.

4.2 Proposition. *Consider $l \in \mathfrak{L}$ and let the loss l_n be defined by* (41). *Then in the Gaussian experiment $\mathbb{E}^0_{\sim,n}$ the minimax risk fulfills*

$$\rho_{\sim,n}(l_n, W_n) \to l(1/2) \text{ as } n \to \infty.$$

Proof. For the lower bound, assume that l is bounded and consider a prior distribution where $f_{(j)}$ are independent $N(0, 1-\delta)$, where $\delta > 0$. By the law of large numbers, this prior concentrates on W_n as $n \to \infty$, so that the Bayes risk is an asymptotic lower bound for $\rho_{\text{-},n}(l_n, W_n)$. The loss $l_n(g, f)$ is subconvex, hence the posterior expectation of f is the Bayes estimator. This Bayes estimator is $\hat{f}_{(j)} = \frac{1-\delta}{1+1-\delta} y_j$, so that the Bayes risk is

$$E\, l(n^{-1} \sum_{j=1}^{n} ((1-\delta) y_j/2 - \delta - f_{(j)})^2) \tag{42}$$

Here $\frac{1-\delta}{2-\delta} y_j - f_{(j)}$ are i. i. d. normal random variables with variance $v_\delta = 2(1-\delta^2)/(2-\delta)^2$, so that (42) converges to $l(v_\delta)$. For $\delta \to 0$ we get $l(v_\delta) \to l(1/2)$.
For attainment of this bound, consider first the case $l \in \mathfrak{L}_b$ and the estimator $\hat{f}_{(j)} = y_j/2$, $j = 1, \ldots, n$. We have for $f \in W_n$

$$n^{-1} \sum_{j=1}^{n} (\hat{f}_{(j)} - f_{(j)})^2 = \frac{1}{4} n^{-1} \sum_{j=1}^{n} \xi_j^2 + \frac{1}{4} n^{-1} \sum_{j=1}^{n} \xi_j f_{(j)} + \frac{1}{4} n^{-1} \sum_{j=1}^{n} f_{(j)}^2 \le \frac{1}{2} + o_p(1).$$

The extension to general $l \in \mathfrak{L}$ takes a few more lines of standard reasoning. □

Pinker's result for Sobolev smoothness classes of functions can be construed as a generalization to infinite dimensional ellipsoids which are "oblique" in the sense of being nonsymmetric in the indices. Let $\varphi_j(x) = \sqrt{2}\cos(2\pi j x)$, $j \ge 1$, $\varphi_j(x) = \sqrt{2}\sin(2\pi j x)$, $j \le -1$, $\varphi_0 \equiv 1$ be the Fourier basis in $L_2(0,1)$, and $f_{(j)} = \langle f, \varphi_j \rangle$ be the Fourier coefficients of a function f. Consider a periodic Sobolev class

$$\tilde{W}_2^\beta(K) = \left\{ f \in L_2(0,1) : \quad \sum_j (2\pi j)^{2\beta} f_{(j)}^2 \le K \right\}.$$

and write $\tilde{W}_2^\beta(1) = \tilde{W}_2^\beta$. We state Pinsker's minimax risk bound in the white noise model, in the variant for bounded l according to Tsybakov (1994). Further discussion of the decison theoretic background can be found in Donoho, Liu and MacGibbon (1990). Let $\mathbb{E}_{\sim,n}$ be the experiment given by observations

$$y_j = f_{(j)} + n^{-1/2} \xi_j, \ j = 1, 2, \ldots \tag{43}$$

where ξ_j are independent standard normal and $f \in \tilde{W}_2^\beta$.

4.3 Proposition (Pinsker, Tsybakov). *Suppose $\beta > 0$ and let $r = \frac{1}{2\beta+1}$. Consider $l \in \mathfrak{L}$ and let the loss l_n be defined by* (39). *Then in the Gaussian experiment $\mathbb{E}_{\sim,n}$ the minimax risk fulfills*

$$\rho_{\sim,n}(l_n, \tilde{W}_2^\beta) \to l(\gamma(\beta)) \text{ as } n \to \infty,$$

where $\gamma(\beta) = (2\beta+1)^r \, (\beta/\pi(\beta+1))^{1-r}$ is the Pinsker constant.

Note that (43) is equivalent to the Gaussian white noise model

$$dy(t) = f(t)dt + n^{-1/2} dW(t), \ t \in [0,1].$$

For application to density estimation, we consider a more general "heteroscedastic" form as in (28)

$$dy(t) = f(t)dt + n^{-1/2} f_0^{1/2}(t) dW(t),\ t \in [0,1] \tag{44}$$

where f_0 is a fixed probability density from the parameter space $\Sigma = \Sigma_{\alpha,M,\epsilon}$ defined in (2). Recall that the distributions of y in (44) were called Q_{2,n,f,f_0} in section 3; let $E_{\sim,n,f,f_0}$ be the respective expectation. Let $\mathbb{E}_{\sim,n}(f_0)$ be the experiment formed by these measures when f_0 is fixed, with parameter space $f \in \tilde{W}_2^\beta$, and let $\rho_{\sim,n,f_0}(\cdot,\cdot)$ be a minimax risk there. Furthermore, we need a localized variant of the risk bound over shrinking uniform neighborhoods. Denote

$$b_0(\tau) = \left\{ f : \int_0^1 f = 0,\ \|f\|_\infty \le \tau \right\}.$$

It turns out that the Pinsker bound holds also in this heteroscedastic case, and in the localized setting. Let $\mathbf{1}$ be the uniform density on $[0,1]$. We restrict ourselves to natural β in order to keep the proof simple (section 11).

4.4 Proposition. *Suppose β is natural and let $r = 1/(2\beta+1)$. Consider $l \in \mathfrak{L}$ and let the loss l_n be defined by* (39).
(i) In the Gaussian experiment $\mathbb{E}_{\sim,n}(\mathbf{1})$, for any sequence: $\tau_n \to 0$, $\tau_n n^{\beta/(2\beta+1)} \to \infty$ we have

$$\liminf_n \rho_{\sim,n,\mathbf{1}}(l_n, \tilde{W}_2^\beta \cap b_0(\tau_n)) \ge l(\gamma(\beta)).$$

(ii) In the Gaussian experiments $\mathbb{E}_{\sim,n}(f_0)$, there is a sequence of estimators $\hat{f}_n^*$, not depending on f_0, such that

$$\limsup_n \sup_{f \in \tilde{W}_2^\beta,\ f_0 \in \Sigma} E_{\sim,n,f,f_0} l_n(\hat{f}_n^*, f) \le l(\gamma(\beta)).$$

We are now ready for application to density estimation. Consider the set of densities

$$\mathcal{W}_\epsilon^\beta = \tilde{W}_2^\beta \cap \mathcal{F}_{\ge\epsilon}.$$

In conjunction with proposition 4.1 this already allows to state a risk convergence in the density model. We first use the local asymptotic equivalence of remark 2.9 for a *lower* asymptotic risk bound. Now have to assume $\beta > 1$, since the Sobolev class $\tilde{W}_2^\beta(K)$ is embedded in a Hölder class $\Lambda^{\beta-1/2}(K')$. Consider the experiment given by (28) with $f \in \Sigma_n(f_0)$ for $f_0 = \mathbf{1}$.

4.5 Proposition. *Suppose β is natural, $\beta > 1$ and let $r = 1/(2\beta+1)$. Consider $l \in \mathfrak{L}_b$ and let the loss l_n be defined by* (39). *Then in the density experiment $\mathbb{E}_{0,n}$ the minimax risk over $\mathcal{W}_\epsilon^\beta$ fulfills*

$$\liminf_n \rho_{0,n}(l_n, \mathcal{W}_\epsilon^\beta) \ge l(\gamma(\beta)).$$

For the converse upper bound we shall invoke the global result of corollary 3.3. Take the model (37) for a choice $N_n = n/\log n$ and look what risk bounds are attainable there.

4.6 Proposition. *Under the conditions of the previous proposition, in the density experiment* $\mathbb{E}_{0,n}$ *the minimax risk over* $\mathcal{W}_\epsilon^\beta$ *fulfills*

$$\limsup_n \ \rho_{0,n}(l_n, \mathcal{W}_\epsilon^\beta) \le l(\gamma(\beta)).$$

We have seen that transferring the Pinsker bound to the density case (cp. the details in section 11) is still somewhat cumbersome; the reason is that a white noise approximation with f as signal is not available in a closed global form. A more direct reasoning is possible for the Hellinger risk of a density, in view of the white noise approximation of theorem 1.1 where $f^{1/2}$ is the signal. This presupposes an adapted a priori class

$$\overline{\mathcal{W}}_\epsilon^\beta = \left\{ f,\ f \in \mathcal{F}_{\ge\epsilon},\ f^{1/2} \in \tilde{W}_2^\beta \right\}.$$

Define a squared Hellinger loss as

$$l_n^H(g,f) = l\left(n^{1-r} \| g^{1/2} - f^{1/2} \|_2^2 \right)$$

4.7 Proposition. *Suppose* β *is natural,* $\beta > 1$ *and let* $r = 1/(2\beta+1)$. *Consider* $l \in \mathfrak{L}_b$ *and let the loss* l_n^H *be defined as above (Hellinger loss). Then in the density experiment* $\mathbb{E}_{0,n}$ *the minimax risk over* $\overline{\mathcal{W}}_\epsilon^\beta$ *fulfills*

$$\rho_{0,n}(l_n^H, \overline{\mathcal{W}}_\epsilon^\beta) \to l\left(2^{2(r-1)} \gamma(\beta) \right) \quad as \quad n \to \infty.$$

Another natural application of asymptotic equivalence is minimax nonparametric hypothesis testing, where a theory of optimal rates and constants is also developing (cp. Ingster (1993)).

Part II: Technical sections

5 Poissonization and Product Structure

For the proof of proposition 2.5 we need some basic concepts from the theory of point processes, see Reiss (1993). A point measure on $(\mathbb{R}, \mathcal{B})$ is a measure $\mu : \mathcal{B} \mapsto [0, \infty]$ of form $\mu = \sum_{i \in I} \mu_{x_i}$, where $I \subset \mathbb{N}$, x_i are points in $\mathbb{R}$ and μ_x is Dirac measure at x. A point process is a random variable on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with values in the space of point measures $\mathbb{M}$ equipped with the appropriate σ-algebra $\mathcal{M}$, see Reiss (1993), p. 6. If $Y = \{y_i,\ i = 1, 2, \ldots\}$ is a sequence of i. i. d. r. v.'s then the random measure $\mu_{0,n} = \sum_{i=1}^n \mu_{y_i}$ is called an empirical point process. More generally if ν is a random natural number independent of Y then $\mu = \sum_{i=1}^\nu \mu_{y_i}$ is a mixed empirical point process. In particular if $\nu = \pi_n$ is Poisson(n) then $\mu_{*,n} = \sum_{i=1}^{\pi_n} \mu_{y_i}$ is a Poisson process which has intensity function nf if y_1 has density f. If f and f_0 are two densities for y_1 such that $P_f \ll P_{f_0}$ and the law of ν is given then it is possible to write down densities for the distributions $\Pi_f := \mathcal{L}(\mu \mid P_f)$ of the mixed empirical point process μ. For the case of the empirical and the Poisson point process ($\nu = n$ or $\nu = \pi_n$) we shall denote these distributions respectively by $\Pi_{0,n,f}$ and $\Pi_{*,n,f}$. For

observations $(\nu, y_i,\ i = 1, \ldots, \nu)$ write the likelihood ratio for hypotheses $(P_f, \mathcal{L}(\nu))$ versus $(P_{f_0}, \mathcal{L}(\nu))$

$$\prod_{i=1}^{\nu}(f/f_0)(y_i) = \exp \int \log(f/f_0) d\mu. \tag{45}$$

This is a function of μ which can be construed as a density of the point process law Π_f on $(\mathbb{M}, \mathcal{M}, \Pi_{f_0})$, or as a likelihood process when f varies. Note that for different $\mathcal{L}(\nu)$ these densities are defined on different probability spaces, since the respective laws Π_{f_0} differ. However let $(\Omega, \mathcal{A}, \mathbb{P}) = ([0,1]^\infty, \mathcal{B}_{[0,1]}^\infty, \lambda^{\otimes\infty})$ where λ is Lebesgue measure on $[0,1]$ and let Y and ν be defined on that space (as independent r. v.'s). Then (45) also describes versions on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ which is common for different $\mathcal{L}(\nu)$. For the case of the empirical and the Poisson point process ($\nu = n$ or $\nu = \pi_n$) we shall denote these likelihood processes respectively by $\Lambda_{0,n}(f, f_0)(\omega)$ and $\Lambda_{*,n}(f, f_0)(\omega)$. The experiments defined by these versions construed as $\mathbb{P}$-densities are then equivalent to the respective point process experiments, for any parameter space. In particular the empirical point process experiment (with laws $\Pi_{0,n,f}$) is equivalent to the original i. i. d. experiment with n observations; $\mu_{0,n} = \sum_{i=1}^n \mu_{y_i}$ is a sufficient statistic.

For our particular parameter space $\Sigma_n(f_0)$ define the Poisson process experiment

$$\mathbb{E}_{*,n}(f_0) = (\mathbb{M}, \mathcal{M}, (\Pi_{*,n,f},\ f \in \Sigma_n(f_0))$$

and recall the definition (8) of the i. i. d. experiment $\mathbb{E}_{0,n}(f_0)$.

5.1 Proposition. *We have*

$$\Delta\left(\mathbb{E}_{0,n}(f_0), \mathbb{E}_{*,n}(f_0)\right) \to 0$$

uniformly over $f_0 \in \Sigma$.

Proof. We use an argument adapted from Le Cam (1985). It suffices to establish that

$$H^2\left(\Lambda_{0,n}(f, f_0), \Lambda_{*,n}(f, f_0)\right) = O(n^{1/2}\gamma_n^2)$$

uniformly over $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma$. With $\nu_{\min} = \min(\pi_n, n)$, $\nu_{\max} = \max(\pi_n, n)$ we get

$$H^2\left(\Lambda_{0,n}(f, f_0), \Lambda_{*,n}(f, f_0)\right) = E_{\mathbb{P}}\left|\prod_{i=1}^{n}(f/f_0)^{1/2}(y_i) - \prod_{i=1}^{\pi_n}(f/f_0)^{1/2}(y_i)\right|^2$$

$$= E_{\mathbb{P}}\prod_{i=1}^{\nu_{\min}}(f/f_0)(y_i)\left|\prod_{i=\nu_{\min}+1}^{\nu_{\max}}(f/f_0)^{1/2}(y_i) - 1\right|^2.$$

Consider first the conditional expectation when π_n is given; since y_i are independent it is

$$E_{\mathbb{P}}\left(\left|\prod_{i=\nu_{\min}+1}^{\nu_{\max}}(f/f_0)^{1/2}(y_i) - 1\right|^2 \mid \pi_n\right).$$

This can be construed as the squared Hellinger distance of two product densities, one of which has $\nu_{\max} - \nu_{\min} = |\pi_n - n|$ factors and the other has as many factors equal to unity. Applying lemma 2.4 we get an upper bound

$$2\sum_{i=\nu_{\min}+1}^{\nu_{\max}} E_{\mathbb{P}}\left(\left|(f/f_0)^{1/2}(y_i) - 1\right|^2 \mid \pi_n\right) \le 2\ |\pi_n - n|\ \gamma_n^2.$$

Taking an expectation and observing $E|\pi_n - n| \le Cn^{1/2}$ completes the proof.□

If μ is a point process and D a measurable set then define the truncated point process

$$\mu_D(B) = \mu(B \cap D),\ B \in \mathcal{B}.$$

Let $\mu_{0,n,D},\ \mu_{*,n,D}$ be truncated empirical and Poisson point process on $[0,1]$, respectively. The following Hellinger distance estimate is due to Falk and Reiss (1992); see also Reiss (1993), theorem 1.4.2:

$$H\left(\mathcal{L}(\mu_{0,n,D} \mid f), \mathcal{L}(\mu_{*,n,D} \mid f)\right) \le \sqrt{3} P_f(D). \tag{46}$$

Proof of proposition 2.5. By the previous proposition it suffices to establish that

$$\Delta(\mathbb{E}_{*,n}(f_0), \bigotimes_{j=1}^{k_n} \mathbb{E}_{0,j,n}(f_0)) \to 0 \tag{47}$$

uniformly over $f_0 \in \Sigma$. In $\mathbb{E}_{0,j,n}(f_0)$ we observe n i. i. d. truncated random variables (18); their empirical point process is a sufficient statistic. Hence μ_{0,n,D_j} (the truncated empirical point process for the original y_i) is a sufficient statistic also; let $\Pi_{0,j,n,f} = \mathcal{L}(\mu_{0,n,D_j} \mid f)$ be the corresponding law. It follows that each $\mathbb{E}_{0,j,n}(f_0)$ is equivalent to an experiment

$$\mathbb{E}^*_{0,j,n}(f_0) = (\mathbb{M}, \mathcal{M}, (\Pi_{0,j,n,f},\ f \in \Sigma_n(f_0))\,.$$

Let $\Pi_{*,j,n,f} = \mathcal{L}(\mu_{*,n,D_j} \mid f)$ be the law of the truncated Poisson point process and

$$\mathbb{E}_{*,j,n}(f_0) = (\mathbb{M}, \mathcal{M}, (\Pi_{*,j,n,f},\ f \in \Sigma_n(f_0))\,;$$

then by the properties of the Poisson process $\mathbb{E}_{*,n}(f_0)$ is equivalent to $\otimes_{j=1}^{k_n} \mathbb{E}_{*,j,n}(f_0)$. It now suffices to show that

$$\Delta(\bigotimes_{j=1}^{k_n} \mathbb{E}_{*,j,n}(f_0), \bigotimes_{j=1}^{k_n} \mathbb{E}^*_{0,j,n}(f_0)) \to 0$$

uniformly over $f_0 \in \Sigma$. From lemma 2.4 and (46) we obtain

$$H^2(\bigotimes_{j=1}^{k_n} \Pi_{*,j,n,f}, \bigotimes_{j=1}^{k_n} \Pi_{0,j,n,f}) \le 2 \sum_{j=1}^{k_n} H^2(\Pi_{*,j,n,f}, \Pi_{0,j,n,f}) \le 6 \sum_{j=1}^{k_n} P_f^2(D_j)$$

$$\le 6 \sup_{1 \le j \le k_n} P_f(D_j).$$

The functions $f \in \Sigma$ are uniformly bounded, in view of the uniform Hölder condition and $\int f = 1$. Hence $P_f(D_j) \to 0$ uniformly in $f \in \Sigma$ and j.□

6 Empirical Processes and Function Classes

From the point process framework we now return to the traditional notion of the empirical process as a normalized and centered random function. However we consider processes indexed by functions. Let $z_i,\ i = 1, \ldots, n$ be i. i. d. uniform random variables on $[0,1]$. Then

$$U_n(f) = n^{1/2} \left(n^{-1} \sum_{i=1}^{n} f(z_i) - \int f \right),\ f \in \mathcal{L}_2[0,1]$$

is the uniform empirical process. The corresponding Brownian bridge process is defined as a centered Gaussian random function $B(f)$, $f \in \mathcal{L}_2[0,1]$ with covariance

$$EB(f)B(g) = \int fg - \left(\int f\right)\left(\int g\right), \ f,g \in \mathcal{L}_2[0,1].$$

For any natural i, consider the subspace of $\mathcal{L}_2[0,1]$ consisting of piecewise constant functions on $[0,1]$ for a partition $[(j-1)2^{-i}, j2^{-i})$, $j = 1,\dots,2^i$. Let $g_{\langle i\rangle}$ be the projection of a function g onto that subspace, and define for natural K

$$q_M(g) = \left(\sum_{i=0}^{K} 2^i \left\|g - g_{\langle i\rangle}\right\|_2^2\right)^{1/2}$$

The following version of a KMT inequality is due to Koltchinskii (1994), theorem 3.5 (specialized to a single element function class $\mathcal{F}$ there and to $K = \log_2 n$)

6.1 Proposition. *There are a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ and numbers C_1, C_2 such that for all n, there are versions $\mathbb{U}_n$ and $\mathbb{B}_n$ of the empirical process and of the Brownian bridge such that for all $g \in \mathcal{L}_2[0,1]$ with $\|g\|_\infty \le 1$ and for all $x, y \ge 0$*

$$\mathbb{P}(n^{1/2}\,|\mathbb{U}_n(g) - \mathbb{B}_n(g)| \ge x + x^{1/2}y^{1/2}(q_{\log_2 n}(g) + 1))$$

$$\le C_1(\exp(-C_2 x) + n\exp(-C_2 y)). \tag{48}$$

To set $q_K(g)$ in relation to a smoothness measure, consider functions $g \in \mathcal{L}_2[0,1]$ satisfying for some C

$$\int_h^{1-h} (g(u+h) - g(u))^2 du \le Ch \text{ for all } h > 0. \tag{49}$$

For a given g, define $\|g\|^2_{H_2^{1/2}}$ as the the infimum of all numbers C for such that (49) holds; it is easy to see that $\|\cdot\|_{H_2^{1/2}}$ is a seminorm. The corresponding space $H_2^{1/2}$ with norm $\|\cdot\|_2 + \|\cdot\|_{H_2^{1/2}}$ coincides with the Besov space $B^{1/2}_{2,\infty}$ on $[0,1]$ (see Nikolskij (1975), 4.3.3, 6.2). Furthermore (cf. Koltchinskii (1994), relation (4.5))

$$q_K^2(g) \le 4K \|g\|^2_{H_2^{1/2}}.$$

Proof of proposition 2.3. If g fulfills $\|g\|_\infty < \infty$ we divide by $\|g\|_\infty$ and apply (48); furthermore we put $y = x + C_2^{-1}\log n$, $x = C_2^{-1}t$ and obtain from (48)

$$2C_1\exp(-t) \ge$$

$$\mathbb{P}(n^{1/2}\,|\mathbb{U}_n(g) - \mathbb{B}_n(g)| \ge \|g\|_\infty x + x^{1/2}y^{1/2}(q_{\log_2 n}(g) + \|g\|_\infty))$$

$$\ge \mathbb{P}(n^{1/2}\,|\mathbb{U}_n(g) - \mathbb{B}_n(g)| \ge \|g\|_\infty x + x^{1/2}y^{1/2}(2\|g\|_{H_2^{1/2}}(\log_2 n)^{1/2} + \|g\|_\infty))$$

$$\ge \mathbb{P}(n^{1/2}\,|\mathbb{U}_n(g) - \mathbb{B}_n(g)| \ge C(\|g\|_\infty + \|g\|_{H_2^{1/2}})(t + \log n)(\log n)^{1/2}).$$

□

6.3 Lemma. *There is a C such that for all $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma$*

$$\|\lambda_{f,f_0}\|_\infty \le C\gamma_n,\ \lambda_{f,f_0} \in \Lambda^\alpha(C).$$

Proof. The first relation is obvious. For the second, note that F_0^{-1} has derivative $1/f(F_0^{-1}(\cdot))$, and since $f \ge \epsilon$, we have $F_0^{-1} \in \Lambda^1(C)$. Now write λ_{f,f_0} as a difference of logarithms and invoke again $f \ge \epsilon$. □

Next we have to consider the likelihood ratio for interval censored observations (18). We shall do this for a generic interval $D \subset [0,1]$ of length k_n^{-1}. We wish to represent the observations via the quantile function F_0^{-1} in the usual fashion; we therefore assume $D = F_0^{-1}(A)$ where $A \subset [0,1]$. Consider a class of intervals, for given $C_1,\ C_2 > 0$,

$$\mathfrak{A}_n = \{A :\ A = [a_1, a_2) \subset [0,1],\ C_1 \le k_n \mathrm{mes}(A) \le C_2\} \tag{50}$$

The assumption $f_0 \in \Sigma$ implies that f_0 is uniformly bounded and bounded away from zero. Hence $\mathrm{mes}(D) = k_n^{-1}$ implies that $A = F_0(D) \in \mathfrak{A}_n$ for all $f_0 \in \Sigma$ and appropriately chosen $C_1,\ C_2$. The technical development will now be carried out uniformly over all intervals $A \in \mathfrak{A}_n$. We shall put $P_f(F_0^{-1}(A)) = p$, $P_{f_0}(F_0^{-1}(A)) = p_0$. The corresponding log-likelihood ratio under f_0, expressed as a function of a uniform $[0,1]$ variable z, is then $\lambda_{f,f_0,A}(z)$, where

$$\lambda_{f,f_0,A}(t) = \chi_A(t) \log \frac{f}{f_0}(F_0^{-1}(t)) + (1 - \chi_A(t)) \log \frac{1-p}{1-p_0}. \tag{51}$$

Since $\lambda_{f,f_0,A}$ has jumps at the endpoints of A, it is not in a Hölder class $\Lambda^\alpha(M)$ but it is in an L_2-Hölder class, so that we can ultimately estimate $\|\lambda_{f,f_0,A}\|_{H_2^{1/2}}$ and apply the KMT-inequality of proposition 2.3. We first need some technical lemmas.

6.4 Lemma. *There is a C such that for all $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma$, $A \in \mathfrak{A}_n$*

$$\sup_{t \in A} |\lambda_{f,f_0,A}(t)| \le C\gamma_n,\ \sup_{t \in A^c} |\lambda_{f,f_0,A}(t)| \le Ck_n^{-1}\gamma_n.$$

Proof. For $t \in A$ we invoke the previous lemma. For $t \in A^c$ we estimate

$$\left|1 - \frac{p}{p_0}\right| \le \frac{\int_D |f - f_0|}{\int_D f_0} \le \frac{\int_D \left|\frac{f}{f_0} - 1\right| f_0}{\int_D f_0} \le \gamma_n.$$

In view of (50) we also have $p_0 \asymp k_n^{-1} \le 1/2$, hence

$$\left|1 - \frac{1-p}{1-p_0}\right| = \frac{p_0}{1-p_0}\left|1 - \frac{p}{p_0}\right| \le Ck_n^{-1}\gamma_n. \tag{52}$$

This implies a similar estimate for $|\log((1-p)/(1-p_0))|$ and thus yields the estimate for $t \in A^c$. □

6.5 Lemma. *There is a constant C such that for all $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma$, $A \in \mathfrak{A}_n$*

$$\int \lambda^2_{f,f_0,A} \le C\, n^{-1}, \; -\int \lambda_{f,f_0,A} \le C\, n^{-1}.$$

Proof. From the previous lemma and (50) we obtain

$$\int \lambda^2_{f,f_0,A} = \int_A \lambda^2_{f,f_0,A} + \int_{A^c} \lambda^2_{f,f_0,A} \le Ck_n^{-1}\gamma_n^2 + Ck_n^{-2}\gamma_n^2 \le Ck_n^{-1}\gamma_n^2, \tag{53}$$

hence in view of (6) and (19)

$$n\int \lambda^2_{f,f_0,A} \le Cnk_n^{-1}\gamma_n^2 \le C.$$

To prove the second relation, define $\varphi(t) = \exp \lambda_{f,f_0,A}(t)$; then $\int \varphi = 1$, and lemma 6.4 implies $|\varphi(t) - 1| \le C\gamma_n$ uniformly. Hence

$$-n\int \lambda_{f,f_0,A} = -n\int \log\varphi \le n\int (1 - \varphi + C(\varphi - 1)^2) = Cn\int (\varphi - 1)^2$$

Here we treat the r. h. s. analogously to (53), using the fact that lemma 6.4 remains true with $\varphi - 1$ in place of λ, so that

$$n\int (\varphi - 1)^2 \le C. \tag{54}$$

□

6.6 Lemma. *There is a C such that for all $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma$, $A \in \mathfrak{A}_n$*

$$\|\lambda_{f,f_0,A}\|_{H_2^{1/2}} \le C\gamma_n.$$

Proof. It suffices to show

$$\int_h^{1-h} (\lambda_{f,f_0,A}(x+h) - \lambda_{f,f_0,A}(x))^2\, dx \le C\gamma_n^2 h \;\text{ for }\; 0 < h < 1/2. \tag{55}$$

Let $A = [a_1, a_2)$ and define $A_{1,h} = [a_1 + h, a_2 - h)$, $A_{2,h} = [a_1 + h, a_2 - h) \cap [h, 1-h]$ (here $A_{1,h}$ is empty for $h > k_n/2$). The integral above over $[h, 1-h]$ will be split into integrals over $A_{1,h}$, $A_{2,h} \setminus A_{1,h}$ and $[h, 1-h] \setminus A_{2,h}$. According to lemma 6.3, $\lambda_{f,f_0,A}$ fulfills a Hölder condition on A, so that

$$\int_{A_{1,h}} (\lambda_{f,f_0,A}(x+h) - \lambda_{f,f_0,A}(x))^2\, dx \le Ch^{2\alpha}k_n^{-1}$$

We have $k_n^{-1} \sim \gamma_n^2(\log n)^4$ in view of (6) and (19), so that $\alpha > 1/2$ implies $Ch^{2\alpha}k_n^{-1} \le Ch\gamma_n^2$. For the second integral, we use the estimate $\|\lambda_{f,f_0,A}\|_\infty \le C\gamma_n$ implied by lemma 6.4, and obtain

$$\int_{A_{2,h}\setminus A_{1,h}} (\lambda_{f,f_0,A}(x+h) - \lambda_{f,f_0,A}(x))^2\, dx \le C\gamma_n^2 h.$$

Finally, note that $\lambda_{f,f_0,A}$ is constant on $[0,1] \setminus A_{2,h}$, so that

$$\int_{[h,1-h]\setminus A_{2,h}} (\lambda_{f,f_0,A}(x+h) - \lambda_{f,f_0,A}(x))^2\, dx = 0.$$

Thus (55) is established.□

7 The local likelihood processes

Consider now the likelihood process for n observations (18) when D_j is replaced by the generic subinterval $D = F_0^{-1}(A)$ with $A \in \mathfrak{A}_n$ from (50). With n i. i. d. uniform $(0,1)$-variables z_i we get an expression for the likelihood process

$$\Lambda_{0,n}(f, f_0, A) = \exp\left\{\sum_{i=1}^{n} \lambda_{f,f_0,A}(z_i)\right\}; \tag{56}$$

for $A = F_0(D_j)$ this is the same as $\Lambda_{0,j,n}(f, f_0)$ as defined after (19). Denote

$$K(f_0 \parallel f, A) = -\int \lambda_{f,f_0,A}(t)dt$$

the pertaining Kullback information number. We assume that $\mathbb{U}_n$ and $\mathbb{B}_n$ are sequences of uniform empirical processes and Brownian bridges which both come from the Hungarian construction of proposition 2.3. We obtain the representation (cp. (14) and proposition 2.6, suppressing the notational distinction of versions)

$$\Lambda_{0,n}(f, f_0, A) = \exp\left\{n^{1/2}\, \mathbb{U}_n(\lambda_{f,f_0,A}) - n\, K(f_0 \parallel f, A)\right\}. \tag{57}$$

The corresponding Gaussian likelihood ratio is (cp. (15))

$$\Lambda_{1,n}(f, f_0, A) = \exp\left\{n^{1/2}\, \mathbb{B}_n(\lambda_{f,f_0,A}) - \frac{n}{2}\mathrm{Var}(\lambda_{f,f_0,A}(Z))\right\}. \tag{58}$$

Consider also an intermediary expression

$$\Lambda_{\#,n}(f, f_0, A) = \exp\left\{n^{1/2}\, \mathbb{B}_n(\lambda_{f,f_0,A}) - n\, K(f_0 \parallel f, A))\right\}.$$

The expression $\Lambda_{\#,n}(f, f_0, A)$ is not normalized to expectation one, but we consider it as the density of a positive measure on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$. The Hellinger distance $H^2(\cdot,\cdot)$ is then naturally extended to these positive measures.

7.1 Lemma. *There is a C such that for all $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma$, $A \in \mathfrak{A}_n$*

$$E_{\mathbb{P}}\left(\Lambda_{i,n}(f, f_0, A)\right)^2 \le C,\ i = 0, 1,\ E_{\mathbb{P}}\left(\Lambda_{\#,n}(f, f_0, A)\right)^2 \le C.$$

Proof. Define (for a uniform $(0,1)$-variable Z)

$$T_{11} = n\, K(f_0 \parallel f, A),\ T_{12} = \frac{n}{2}\mathrm{Var}(\lambda_{f,f_0,A}(Z)), \tag{59}$$

$$T_{21} = n^{1/2}\, \mathbb{U}_n(\lambda_{f,f_0,A}),\ T_{22} = n^{1/2}\, \mathbb{B}_n(\lambda_{f,f_0,A}). \tag{60}$$

Since T_{22} is a zero mean Gaussian r. v., we have

$$E_{\mathbb{P}} \exp(2T_{22}) = \exp\left(4T_{12}\right).$$

Hence

$$E_{\mathbb{P}}\Lambda_{1,n}^2 = E_{\mathbb{P}} \exp\left(2\left(T_{22} - T_{12}\right)\right) = \exp\left(2T_{12}\right) \le \exp\left(n \int \lambda_{f,f_0,A}^2\right).$$

Now from lemma 6.5 we obtain the assertion for $i = 1$. For the case $i = 0$, we get from (56)

$$E_{\mathbb{P}}\Lambda_{0,n}^2 = E_{\mathbb{P}} \exp\left\{ 2\sum_{i=1}^{n} \lambda_{f,f_0,A}(z_i) \right\} = \left(E \exp\left(2\lambda_{f,f_0,A}(Z)\right)\right)^n .$$

Now we have for $\varphi(t) = \exp \lambda_{f,f_0,A}(t)$

$$E \exp 2\lambda_{f,f_0,A}(Z) = \int (\varphi(t))^2 dt = 1 + \int \left(\varphi(t) - 1\right)^2 dt \leq 1 + Cn^{-1}$$

as a consequence of (54). Hence

$$E_{\mathbb{P}}\Lambda_{0,n}^2 \leq \left(1 + Cn^{-1}\right)^n \leq 2 \exp C$$

so that the lemma is established for $i = 0$. Finally, to treat $E_{\mathbb{P}}\Lambda_{\#,n}^2$, observe that lemma 6.5 implies that T_{11} and T_{12} are uniformly bounded. Hence

$$E_{\mathbb{P}}\Lambda_{\#,n}^2 = E_{\mathbb{P}}\Lambda_{1,n}^2 \exp\left(2(T_{12} - T_{11})\right) \leq C.$$

□.

The next lemma is the key technical step, bringing in the Hungarian construction estimate of proposition 2.3.

7.2 Lemma. *There is a C such that for all $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma$, $A \in \mathfrak{A}_n$*

$$H\left(\Lambda_{0,n}(f, f_0, A), \Lambda_{\#,n}(f, f_0, A)\right) \leq C\, \gamma_n (\log n)^{3/2}.$$

Proof. Define

$$T_0 = n^{1/2}(\mathbb{B}_n - \mathbb{U}_n)(\lambda_{f,f_0,A}).$$

Combining proposition 2.3 with lemmas 6.4 and 6.6, we obtain

$$\mathbb{P}(|T_0| \geq C\gamma_n(t + \log n) \log^{1/2} n) \leq C \exp(-t).$$

Put $t = t_n = 4 \log n$ and for the above C

$$u_n = 5C\gamma_n \log^{3/2} n.$$

For an event

$$B = B_{f,f_0,A} = \{\omega :\ |T_0| \leq u_n\}$$

we obtain an estimate

$$\mathbb{P}(B^c) \leq Cn^{-4}. \tag{61}$$

To treat $H^2(\Lambda_{0,n}, \Lambda_{\#,n})$, split the expectation there into $E_{\mathbb{P}}\chi_B(\cdot)$ and $E_{\mathbb{P}}\chi_{B^c}(\cdot)$, and observe

$$\begin{aligned} E_{\mathbb{P}}\chi_{B^c}(\Lambda_{0,n}^{1/2} - \Lambda_{\#,n}^{1/2})^2 &\leq 2E_{\mathbb{P}}\chi_{B^c}(\Lambda_{0,n} + \Lambda_{\#,n}) \\ &\leq 2\left(\mathbb{P}(B^c) 2E_{\mathbb{P}}(\Lambda_{0,n}^2 + \Lambda_{\#,n}^2)\right)^{1/2}. \end{aligned}$$

According to the previous lemma $E_{\mathbb{P}}(\Lambda_{0,n}^2 + \Lambda_{\#,n}^2)$ is uniformly bounded, so that (61) implies

$$E_{\mathbb{P}}\chi_{B^c}(\Lambda_{0,n}^{1/2} - \Lambda_{\#,n}^{1/2})^2 \leq Cn^{-2}. \tag{62}$$

For the other part, observe that on $\omega \in B$, in view of $u_n = o(1)$,

$$|1 - \exp(T_0/2)| \leq Cu_n,$$

so that on $\omega \in B$

$$(\Lambda_{0,n}^{1/2} - \Lambda_{\#,n}^{1/2})^2 = (1 - \exp(T_0/2))^2 \Lambda_{0,n} \leq Cu_n^2 \Lambda_{0,n}.$$

Since $E_{\mathbb{P}}\Lambda_{0,n} = 1$, we obtain

$$E_{\mathbb{P}}\chi_B(\Lambda_{0,n}^{1/2} - \Lambda_{\#,n}^{1/2})^2 \leq Cu_n^2.$$

This completes the proof in view of (62) and $n^{-2} = o(u_n^2)$.□

7.3 Lemma. *For all $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma$, $A \in \mathfrak{A}_n$*

$$H(\Lambda_{0,n}(f, f_0, A), \Lambda_{1,n}(f, f_0, A)) \leq 2H(\Lambda_{0,n}(f, f_0, A), \Lambda_{\#,n}(f, f_0, A)).$$

Proof. Consider the space of random variables $L_2(\Omega, \mathcal{A}, \mathbb{P})$ and note that $H(\Lambda_{\#,n}, \Lambda_{1,n})$ is the distance of $\Lambda_{\#,n}^{1/2}$ and $\Lambda_{1,n}^{1/2}$ in that space. Furthermore

$$\Lambda_{1,n}^{1/2} = \Lambda_{\#,n}^{1/2} \left(E_{\mathbb{P}}\Lambda_{\#,n}\right)^{-1/2}.$$

is the element of the unit sphere of $L_2(\Omega, \mathcal{A}, \mathbb{P})$ closest to $\Lambda_{\#,n}^{1/2}$. Since $\Lambda_{0,n}^{1/2}$ is on the unit sphere, we have

$$H(\Lambda_{\#,n}, \Lambda_{1,n}) \leq H(\Lambda_{\#,n}, \Lambda_{0,n})$$

and therefore

$$H(\Lambda_{0,n}, \Lambda_{1,n}) \leq H(\Lambda_{0,n}, \Lambda_{\#,n}) + H(\Lambda_{\#,n}, \Lambda_{1,n}) \leq 2H(\Lambda_{0,n}, \Lambda_{\#,n}).$$

□

Let now $A = A_j = F_0(D_j)$ and consider also the likelihood process $\Lambda_{1,j,n}(f, f_0)$ of the Gaussian experiment $\mathbb{E}_{1,j,n}(f_0)$ of (20). Remind that this differs from $\Lambda_{1,n}(f, f_0, A_j)$ (cp. (58) and (51)). We consider versions of both likelihood processes which are functions of the Brownian bridge version $\mathbb{B}$.

7.4 Lemma. *There is a C such that for all $f \in \Sigma_n(f_0)$, $f_0 \in \Sigma$ and $j = 1, \ldots, k_n$*

$$H(\Lambda_{1,n}(f, f_0, A_j), \Lambda_{1,j,n}(f, f_0)) \leq C\gamma_n.$$

Proof. The likelihood process $\Lambda_{1,n}(f, f_0, A_j)$ is $\Lambda_{1,n}(f, f_0)$ from (15) with λ_{f,f_0} replaced by λ_{f,f_0,A_j}, so it corresponds to a Gaussian model

$$dy(t) = (\lambda_{f,f_0,A_j}(t) + K(f_0 \| f, A_j))dt + n^{-1/2}dW(t), \ t \in [0,1]$$

with $f \in \Sigma_n(f_0)$ (cp. (16)). Moreover $\Lambda_{1,j,n}(f,f_0)$ corresponds to the Gaussian model (20). Hence the distance $H(\cdot,\cdot)$ between the likelihood processes on $(\Omega,\mathcal{A},\mathbb{P})$ equals the Hellinger distance between the two respective shifted Wiener measures. We may apply formula (23), putting

$$g_1 = \lambda_{f,f_0,A_j} - \int \lambda_{f,f_0,A_j},\ g_2 = \chi_{A_j}\lambda_{f,f_0} - \int_{A_j} \lambda_{f,f_0}.$$

We obtain in accordance with (13) and (51)

$$\|g_1 - g_2\|_2^2 = \left\| \chi_{A_j^c}\lambda_{f,f_0,A_j} - \int_{A_j^c} \lambda_{f,f_0,A_j} \right\|_2^2 = p_0(1-p_0)\log^2 \frac{1-p}{1-p_0}$$

where $p = P_f(D_j)$, $p_0 = P_{f_0}(D_j)$. Using $p_0 \leq Ck_n^{-1}$ and (52) we find

$$\|g_1 - g_2\|_2^2 \leq Ck_n^{-3}\gamma_n^2.$$

By (23) the squared Hellinger distance is

$$2\left(1 - \exp\left\{-\frac{n}{8}\|g_1 - g_2\|_2^2\right\}\right) \leq 2\left(1 - \exp\left\{-Cnk_n^{-3}\gamma_n^2\right\}\right)$$

and the lemma follows from $nk_n^{-3} = o(1)$.□

Proof of proposition 2.6. Consider $\Lambda_{0,n}(f,f_0,A)$ for $A = A_j$ and identify this to $\Lambda^*_{0,j,n}(f,f_0)$. Identify $\Lambda_{1,j,n}(f,f_0)$ of lemma 5.3.4 to $\Lambda^*_{1,j,n}(f,f_0)$. The result then follows from lemmas 7.2-7.4.□

8 Further local approximations

Define functions

$$\begin{aligned} \lambda_{1,f,f_0} &= \lambda_{f,f_0} + K(f_0\|f),\ \lambda_{2,f,f_0} = (f/f_0 - 1)\circ F_0^{-1}, \\ \lambda_{3,f,f_0} &= 2\left((f/f_0)^{1/2} - 1\right)\circ F_0^{-1} \end{aligned}$$

and experiments $\mathbb{E}^{\#}_{i,n}(f_0)$ given by observations

$$dy(t) = \lambda_{i,f,f_0}(t)dt + n^{-1/2}dW(t),\ t \in [0,1], \tag{63}$$

and parameter space $f \in \Sigma_n(f_0)$, for $i = 1,2,3$. We have seen that $\mathbb{E}^{\#}_{1,n}(f_0) = \mathbb{E}_{1,n}(f_0)$ (cp. (16)).

8.1 Lemma. We have

$$\Delta\left(\mathbb{E}^{\#}_{i,n}(f_0), \mathbb{E}_{i,n}(f_0)\right) = 0,\ i = 1,2,3.$$

Proof. The likelihood process for $\mathbb{E}^{\#}_{i,n}(f_0)$ is

$$\Lambda_{i,n}(f,f_0) = \exp\left\{n\int \lambda_{i,f,f_0}\frac{1}{\sqrt{n}}\,dW - \frac{n}{2}\|\lambda_{i,f,f_0}\|_2^2\right\},\ i = 1,2,3.$$

Define a process

$$W^*(t) = \int_0^t f_0^{-1/2} d(W \circ F_0).$$

This is a centered Gaussian process with independent increments and variance at t given by $\int_0^t f_0^{-1} dF_0 = t$. Hence W^* is a Wiener process, and we have for every continuous g on $[0,1]$

$$\int g f_0^{1/2} dW^* = \int g d(W \circ F_0).$$

Utilizing W^* in (24), we get a likelihood process for $\mathbb{E}_{2,n}(f_0)$

$$\begin{aligned} & \exp\left\{ n \int (f - f_0) f_0^{-1} n^{-1/2} f_0^{1/2} dW^* - \frac{n}{2} \int (f - f_0)^2 f_0^{-1} \right\} \\ = \; & \exp\left\{ n \int (f/f_0 - 1)\, n^{-1/2} d(W \circ F_0) - \frac{n}{2} \int (f/f_0 - 1)^2 \, dF_0 \right\} = \Lambda_{2,n}(f, f_0). \end{aligned}$$

Similarly for $\mathbb{E}_{3,n}(f_0)$ we obtain a likelihood process

$$\begin{aligned} & \exp\left\{ 4n \int (f^{1/2} - f_0^{1/2}) \frac{1}{2} n^{-1/2} dW^* - \frac{4n}{2} \int (f^{1/2} - f_0^{1/2})^2 \right\} \\ = \; & \exp\left\{ 2n \int \left((f/f_0)^{1/2} - 1 \right) n^{-1/2} d(W \circ F_0) - \frac{4n}{2} \int \left((f/f_0)^{1/2} - 1 \right)^2 dF_0 \right\} \\ = \; & \Lambda_{3,n}(f, f_0). \end{aligned}$$

□

Proof of theorem 2.8. It now remains to apply formula (23) to the measures given by (63) when $f \in \Sigma_n(f_0)$. We have to prove

$$\sup_{f \in \Sigma_n(f_0)} \|\lambda_{1,f,f_0} - \lambda_{i,f,f_0}\|_2^2 = o(n^{-1}) \tag{64}$$

for $i = 2, 3$, uniformly over $f_0 \in \Sigma$. Using the expansion

$$\log x = \log(1 + x - 1) = x - 1 - \frac{1}{2}(x-1)^2 + o((x-1)^2) \tag{65}$$

and putting $x = (f/f_0) \circ F_0^{-1}(t)$, we note that for $f \in \Sigma_n(f_0)$

$$\lambda_{f,f_0}(t) = \lambda_{2,f,f_0}(t) + O(\gamma_n^2) \tag{66}$$

uniformly. Since $\int \lambda_{2,f,f_0} = 0$, we obtain

$$K(f_0 \| f) = \int (\lambda_{2,f,f_0} - \lambda_{f,f_0}) \leq \|\lambda_{2,f,f_0} - \lambda_{f,f_0}\|_2 = O(\gamma_n^2). \tag{67}$$

Now (66) and (67) imply

$$\|\lambda_{f,f_0} + K(f_0 \| f) - \lambda_{2,f,f_0}\|_2^2 = O(\gamma_n^4) = O(n^{-1} (\log n)^{-4})$$

which proves (64) for $i = 2$. For $i = 3$, note first that for $f \in \Sigma_n(f_0)$ we have

$$\left\| (f/f_0)^{1/2} - 1 \right\|_\infty = O(\gamma_n),$$

and use (65) with $x = (f/f_0)^{1/2} \circ F_0^{-1}(t)$ to obtain

$$\lambda_{f,f_0}(t) = 2\log(f/f_0)^{1/2} \circ F_0^{-1}(t) = \lambda_{3,f,f_0}(t) + O(\gamma_n^2) \tag{68}$$

uniformly. Now (68) and (67) imply (64) for $i = 3$. □

9 The preliminary estimator

Consider first a histogram estimator based on the whole sample. Let $\psi_{n,\kappa} = (\kappa \log n/n)^{\alpha/(2\alpha+1)}$ for a $\kappa > 0$ and $s_n = [\psi_{n,\kappa}^{-1/\alpha}] + 1$. Define intervals $J_{j,n} = s_n^{-1}[j - 1, j)$, $j = 1, \ldots, s_n$ and let $\bar{F}_n$ be the empirical distribution function of $y_1, \ldots, y_n$. Define an estimator

$$\tilde{f}_n = s_n \sum_{j=1}^{s_n} \chi_{J_{j,n}} \int \chi_{J_{j,n}} d\bar{F}_n.$$

9.1 Lemma. *In the experiment $\mathbb{E}_{0,n}$ there is a κ such that*

$$\sup_{f \in \Sigma} P_{n,f} \left(\left\| \tilde{f}_n - f \right\|_\infty \geq \kappa\, \psi_{n,\kappa} \right) \to 0.$$

Proof. Consider the usual decomposition

$$\left\| \tilde{f}_n - f \right\|_\infty \leq \left\| \tilde{f}_n - E\tilde{f}_n \right\|_\infty + \left\| E\tilde{f}_n - f \right\|_\infty.$$

Note that for $t \in J_{j,n}$

$$\begin{aligned} \left| E\tilde{f}_n(t) - f(t) \right| &= \left| f(t) - s_n \int_{J_{j,n}} f(u)du \right| \leq s_n \int_{J_{j,n}} |f(t) - f(u)|\, du \\ &\leq M s_n \int_{J_{j,n}} |t-u|^\alpha\, du \leq M s_n^{-\alpha} \leq M\psi_{n,\kappa}, \end{aligned}$$

so that

$$\left\| E\tilde{f}_n - f \right\|_\infty \leq M\psi_{n,\kappa}.$$

For the variance part, write for $t \in J_{j,n}$ and observations y_i having density f

$$\begin{aligned} \tilde{f}_n(t) - E\tilde{f}_n(t) &= s_n \int \chi_{J_{j,n}} d(\bar{F}_n - F) = s_n n^{-1} \sum_{i=1}^n \eta_{ij}, \\ \text{where } \eta_{ij} &= \chi_{J_{j,n}}(y_i) - P_f(J_{j,n}). \end{aligned}$$

Then $|\eta_{ij}| \leq 1$ and using notation $v_n = \sum_{i=1}^n \mathrm{Var}(\eta_{ij})$ consider Bernstein's inequality

$$P_{n,f}\left(\left| \sum_{i=1}^n \eta_{ij} \right| \geq t \right) \leq 2\exp\left(-\frac{1}{2} t^2/(v_n + t/3) \right).$$

It is easy to verify that the quantity

$$\mu_\Sigma = \sup_{f\in\Sigma} \|f\|_\infty \tag{69}$$

is finite; this is a consequence of Hölder continuity in conjunction with $\int |f| = 1$. We find

$$v_n = nP_f(J_{j,n})(1 - P_f(J_{j,n})) \le ns_n^{-1}\mu_\Sigma.$$

Putting $t = \psi_{n,\kappa}s_n^{-1}n$, we obtain $v_n + t/3 \le 2ns_n^{-1}\mu_\Sigma$ for large n and

$$P_{n,f}\left(s_n n^{-1}\left|\sum_{i=1}^n \eta_{ij}\right| \ge \psi_{n,\kappa}\right) \le 2\exp\left(-\psi_{n,\kappa}^2 s_n^{-1}n(4\mu_\Sigma)^{-1}\right) \le 3\exp\left(-\kappa(\log n)(4\mu_\Sigma)^{-1}\right).$$

Consequently for $\kappa \ge 4\mu_\Sigma$

$$\begin{aligned} P_{n,f}\left(\left\|\tilde f_n - E\tilde f_n\right\|_\infty \ge \psi_{n,\kappa}\right) &\le \sum_{j=1}^{s_n} P_{n,f}\left(s_n n^{-1}\left|\sum_{i=1}^n \eta_{ij}\right| \ge \psi_{n,\kappa}\right) \\ &\le 3s_n n^{-1} \to 0. \end{aligned}$$

For $\kappa \ge \max(4\mu_\Sigma, 2M, 2)$ we obtain the lemma.□

Proof of lemma 3.1. Consider the estimator applied to a sample fraction y_i, $i = 1, \dots, N_n$; call it $\tilde f_{N_n}$. Then, since $\alpha > 1/2$,

$$\psi_{N_n} = \left(N_n^{-1}\kappa \log N_n\right)^{\alpha/(2\alpha+1)} \le \left(n^{-1}\kappa\log(n/2)\log n\right)^{\alpha/(2\alpha+1)} = o(\gamma_n).$$

This immediately implies

$$\sup_{f\in\Sigma} P_{n,f}\left(\sup_{t\in[0,1]}\left|f(t) - \tilde f_{N_n}(t)\right| > c\gamma_n\right) \to 0, \text{ for all } c > 0. \tag{70}$$

Note that the set Σ is compact in the uniform metric: indeed it is equicontinuous and uniformly bounded according to (69), so compactness is implied by the Arzela-Ascoli theorem. Now cover Σ by a finite set of uniform γ_n-balls with centers in Σ and define $\Sigma_{0,n}$ be the set of the centers. Define $\hat f_n$ as the element in $\Sigma_{0,n}$ closest to $\tilde f_{N_n}$ (or in case of nonuniqueness, select an element measurably). Analogously, for $f \in \Sigma$ select a closest element $g_f \in \Sigma_{0,n}$. Then we have

$$\begin{aligned} \left\|\hat f_n - f\right\|_\infty &\le \left\|\hat f_n - \tilde f_{N_n}\right\|_\infty + \left\|\tilde f_{N_n} - f\right\|_\infty \\ &\le \left\|g_f - \tilde f_{N_n}\right\|_\infty + \left\|\tilde f_{N_n} - f\right\|_\infty \\ &\le \|g_f - f\|_\infty + 2\left\|\tilde f_{N_n} - f\right\|_\infty \le 2\left\|\tilde f_{N_n} - f\right\|_\infty + \gamma_n. \end{aligned}$$

Hence $\hat f_n$ also satisfies (70), and it takes values in the finite set $\Sigma_{0,n} \subset \Sigma$. From this we obtain immediately

$$\sup_{f\in\Sigma} P_{n,f}\left(\sup_{t\in[0,1]}\left|f(t)/\tilde f_{N_n}(t) - 1\right| > \gamma_n\right) \to 0$$

in view of the uniform bound $f(t) \geq \epsilon$ for $f \in \Sigma$. $\square$

For lemma 3.4, we first consider estimation of the signal (rather than its root) in the white noise model. Let again $\psi_{n,\kappa} = (\kappa \log n/n)^{\alpha/(2\alpha+1)}$.

9.3 Lemma. *Consider an experiment given by observations*

$$dy(t) = g(t)dt + n^{-1/2}dW(t), \ t \in [0,1] \tag{71}$$

with $g \in \Lambda^\alpha(M)$. There one can find an estimator $\tilde{g}_n$ and a κ such that

$$\sup_{g \in \Lambda^\alpha(M)} P_{n,g}\left(\|\tilde{g}_n - g\|_\infty \geq \kappa\, \psi_{n,\kappa}\right) \to 0.$$

The proof could be analogous to lemma 9.1, with simplifications due to Gaussianity. Alternatively, we may refer to theorem C in Donoho (1994) where sharper results (optimal constants) are obtained.

Proof of lemma 3.4. If $g = f^{1/2}$ with $f \in \Sigma$ then since $f \in \mathcal{F}_{\geq \epsilon}$

$$\left|f^{1/2}(t) - f^{1/2}(u)\right| \leq \epsilon^{-1/2}\left|f(t) - f(u)\right|$$

so we obtain $g \in \Lambda^\alpha(\epsilon^{-1/2}M)$. Also, by the previous argument we may assume that $\tilde{g}_n$ takes values in a finite subset of $\{f^{1/2} : \ f \in \Sigma\}$. On the other hand, if $\check{f}_n = \tilde{g}_n^2$ then

$$\left|\check{f}_n(t) - f(t)\right| \leq \left|\tilde{g}_n(t) + g(t)\right| \left|\tilde{g}_n(t) - g(t)\right|.$$

Since both $\tilde{g}_n$ and g are in $\{f^{1/2} : \ f \in \Sigma\}$ they are uniformly bounded by $\mu_\Sigma^{1/2}$ (cf. (69)), so that for some κ

$$\sup_{f \in \Sigma} P_{n,f}\left(\left\|\check{f}_n - f\right\|_\infty \geq \kappa\, \psi_n\right) \to 0.$$

Finally assume that $\check{f}_n$ is based on observations with noise intensity $(n - N_n)^{-1/2}$ instead of $n^{-1/2}$, i. e. on (38). Then $(n - N_n)^{-1/2} \leq (n/2)^{-1/2}$ so that attainable rates are not worse. As in lemma 3.1 we now infer that the estimator $\check{f}_n$ based on (38) fulfills (30).$\square$

10 Experiments and globalization

We collect some basic facts about experiments and deficiencies following Strasser (1985) ([S] henceforth). Let $\mathbb{E}_1 = (\Omega_1, \mathcal{A}_1, (P_{1,\vartheta}, \vartheta \in \Theta))$ be an experiment and let $L(\mathbb{E}_1)$ be the corresponding L-space (see [S] 41.4); $L(\mathbb{E}_1)$ is a certain subspace of the set of signed measures on $(\Omega_1, \mathcal{A}_1)$ which is a Banach lattice under the variational norm $\|\cdot\|$. Let $\mathbb{E}_2 = (\Omega_2, \mathcal{A}_2, (P_{2,\vartheta}, \ \vartheta \in \Theta))$ be another experiment with the same parameter set Θ with L-space $L(\mathbb{E}_2)$. A transition from $L(\mathbb{E}_1)$ to $L(\mathbb{E}_2)$ is a positive linear map with norm one (i. e. a linear map $M : L(\mathbb{E}_1) \longmapsto L(\mathbb{E}_2)$ such that for $\sigma \in \mathbb{E}_1$, $\sigma \geq 0$ one has $M\sigma \geq 0$ and $\|M\sigma\| = \|\sigma\|$, cp. [S] 55.2). Every Markov kernel $K : \Omega_1 \times \mathcal{A}_2 \longmapsto [0,1]$ defines a transition. For the definition of the deficiency $\delta(\mathbb{E}_1, \mathbb{E}_2)$ of $\mathbb{E}_1$ with respect to $\mathbb{E}_2$ via decision problems see [S] section 59. An equivalent characterization is ([S] 59.6)

$$\delta(\mathbb{E}_1, \mathbb{E}_2) = \inf_M \sup_{\vartheta \in \Theta} \|MP_{1,\vartheta} - P_{2,\vartheta}\| \tag{72}$$

where the infimum extends over all transitions from $L(\mathbb{E}_1)$ to $L(\mathbb{E}_2)$. The two sided deficiency is

$$\Delta(\mathbb{E}_1, \mathbb{E}_2) = \max(\delta(\mathbb{E}_1, \mathbb{E}_2), \delta(\mathbb{E}_2, \mathbb{E}_1)).$$

This defines a pseudodistance on the set of all experiments with parameter space Θ; in particular, the triangle inequality holds [S] 59.2). $\mathbb{E}_1$ and $\mathbb{E}_2$ are called equivalent (or of the same type) if $\Delta(\mathbb{E}_1, \mathbb{E}_2) = 0$.
We are interested in conditions under which every transition is given by a Markov kernel. [S] 55.6 (3) gives it for the case that $\mathbb{E}_1$ is dominated and Ω_2 is a locally compact space with countable base and $\mathcal{A}_2$ is its Borel σ-algebra. But spaces like $C[0,1]$ are not locally compact, so we would like to have the result for a complete separable metric (Polish) space instead. We briefly complete the argument.

10.1 Definition. *An experiment $\mathbb{E} = (\Omega, \mathcal{A}, (P_\vartheta, \vartheta \in \Theta))$ is called Polish if Ω is a Polish space and $\mathcal{A}$ is the pertaining Borel σ-algebra.*

10.2 Proposition. *Suppose that $\mathbb{E}_1$ is a dominated experiment and $\mathbb{E}_2$ is Polish. Then every transition from $L(\mathbb{E}_1)$ to $L(\mathbb{E}_2)$ is given by a Markov kernel.*

Proof. It is well known that $(\Omega_2, \mathcal{A}_2)$ is Borel isomorphic to a subset of the unit interval (Dudley (1989), lemma 13.1.3, Parthasarathy (1980), Proposition 25.6). This means that there is a one-to-one function φ from Ω_2 onto a Borel subset S of the unit interval such that φ and φ^{-1} are both measurable. It is clear that $\mathbb{E}_2$ is then equivalent to an experiment $\mathbb{E}_2^*$ given on the measurable space $(S, \mathcal{B}_S)$, and this equivalence is realized by Markov kernel transitions given by the mappings φ and φ^{-1}. Thus it suffices to prove the theorem for $\mathbb{E}_2 = \mathbb{E}_2^*$. We now refer to remark 5.5.6 (3) in [S].□

For the proof of theorem 3.2 we formulate a lemma in an abstract framework. Let $\mathbb{X} = (X, \mathcal{X}, (P_\vartheta, \vartheta \in \Theta))$ be an experiment. Suppose also that there are a system of subsets $\Theta(\phi) \subset \Theta$, $\phi \in \Theta$ and experiments

$$\mathbb{F}_i(\phi) = (\Omega_i, \mathcal{A}_i, (Q_{i,\vartheta,\phi},\ \vartheta \in \Theta(\phi))),\ i = 1, 2,\ \phi \in \Theta.$$

Suppose further that there is a finite subset of $\Theta_0 \subset \Theta$ and an estimator $\hat{\phi} : (X, \mathcal{X}) \mapsto (\Theta_0, 2^{\Theta_0})$ and form Markov kernels

$$Q_{i,\vartheta}(x, A') = Q_{i,\vartheta,\hat{\phi}(x)}(A'),\ x \in X,\ A' \in \mathcal{A}_i,\ i = 1, 2.$$

Let $(\bar{X}_i, \bar{\mathcal{X}}_i) = (X \times \Omega_i, \mathcal{X} \times \mathcal{A}_i)$ be a product measurable space. For any Markov kernel $K : X \times \mathcal{A}_i \mapsto [0,1]$ and a measure $\mu \mid \mathcal{X}$ we shall form the usual composed measure $\mu \otimes K \mid \bar{\mathcal{X}}_i$. Define measures $P_{i,\vartheta} \mid \bar{\mathcal{X}}_i = P_\vartheta \otimes Q_{i,\vartheta} \mid \bar{\mathcal{X}}_i$ and experiments $\mathbb{F}_i = (\bar{X}_i, \bar{\mathcal{X}}_i, (P_{i,\vartheta},\ \vartheta \in \Theta))$, $i = 1, 2$.

10.3 Lemma. *Suppose that for all $\phi \in \Theta$ the experiments $\mathbb{F}_i(\phi)$, $i = 1, 2$ are Polish and dominated, and*

$$\sup_{\phi \in \Theta} \Delta(\mathbb{F}_1(\phi), \mathbb{F}_2(\phi)) \leq \epsilon. \tag{73}$$

Suppose also that the estimator $\hat{\phi}$ with values in Θ_0 fulfills

$$\inf_{\vartheta \in \Theta} P_\vartheta(\vartheta \in \Theta(\hat{\phi})) \geq 1 - \epsilon. \tag{74}$$

Then

$$\Delta(\mathbb{F}_1, \mathbb{F}_2) \leq 3\epsilon.$$

Proof. Observe that since Θ_0 is finite and $\hat{\phi}$ is 2^{Θ_0}- measurable, the set $V_\vartheta = \{x : \vartheta \in \Theta(\hat{\phi}(x))\}$ is in $\mathcal{X}$. In accordance with proposition 10.2, let $K_\phi(\omega_2, \cdot)$ be a Markov kernel realizing

$$\delta(\mathbb{F}_1(\phi), \mathbb{F}_2(\phi)) = \sup_{\vartheta \in \Theta(\phi)} \|Q_{2,\vartheta,\phi} - K_\phi Q_{1,\vartheta,\phi}\| \leq \epsilon$$

and define

$$M(\bar{x}, A) = \int_{\Omega_2} \chi_A(x, \omega_2) K_{\hat{\phi}(x)}(\omega_1, d\omega_2),\ \bar{x} = (x, \omega_1) \in \bar{X}_1,\ A \in \bar{\mathcal{X}}_2.$$

It is easy to see that M is a Markov kernel. Indeed by standard arguments this claim is reduced to the measurability of $K_{\hat{\phi}(x)}(\omega_1, A')$ in $\bar{x} = (x, \omega_1)$ for given $A' \in \mathcal{A}_2$, which again follows from the properties of $\hat{\phi}$. Now we have for $A \in \bar{\mathcal{X}}_2$

$$MP_{1,\vartheta}(A) = \int_X \int_{\Omega_1} M(x, \omega_1, A) Q_{1,\vartheta}(x, d\omega_1) P_\vartheta(dx)$$

$$= \int_X \int_{\Omega_2} \chi_A(x, \omega_2) (K_{\hat{\phi}(x)} Q_{1,\vartheta,\hat{\phi}(x)})(d\omega_2) P_\vartheta(dx).$$

Hence

$$|P_{2,\vartheta}(A) - MP_{1,\vartheta}(A)| \leq 2P_\vartheta(V_\vartheta^c)$$

$$+ \int_{V_\vartheta} \left| \int_{\Omega_2} \chi_A(x, \omega_2) (K_{\hat{\phi}(x)} Q_{1,\vartheta,\hat{\phi}(x)} - Q_{2,\vartheta,\hat{\phi}(x)})(d\omega_2) \right| P_\vartheta(dx)$$

$$\leq 2P_\vartheta(V_\vartheta^c) + \sup_{\phi \in \Theta_0} \sup_{\vartheta \in \Theta(\phi)} \|K_\phi Q_{1,\vartheta,\phi} - Q_{2,\vartheta,\phi}\| \leq 3\epsilon$$

and we obtain

$$\delta(\mathbb{F}_1, \mathbb{F}_2) \leq \sup_{\vartheta \in \Theta} \|P_{2,\vartheta} - MP_{1,\vartheta}\| \leq 3\epsilon.$$

The argument for $\delta(\mathbb{F}_2, \mathbb{F}_1)$ is symmetric.□

Proof of theorem 3.2. In the previous lemma we put $\vartheta = f$, $\phi = f_0$, $\Theta = \Sigma$, $\Theta(\phi) = \Sigma_n(f_0)$ and identify the experiment $\mathbb{X}$ to the one given by the sample fraction $y_1, \ldots, y_{N_n}$ (which may be written $\mathbb{E}_{0,N_n}$). Furthermore $\mathbb{F}_1(\phi)$ is given by the second sample fraction with f restricted to a neighborhood $\Sigma_n(f_0)$ (which may be written $\mathbb{E}_{0,n-N_n}(f_0)$, cp. (8)). $\mathbb{F}_2(\phi)$) is given by one of the three local experiments (27), (28), (29) in remark 2.9 (we have seen that those are asymptotically or exactly equivalent to the respective $\mathbb{E}_{j,n}(f_0), j = 1, 2, 3$ from theorem 2.8). Note that both $\mathbb{F}_i(\phi)$), $i = 1, 2$ are then Polish and dominated; in particular, $C_{[0,1]}$ is a Polish space (see Dudley (1989), Corollary 11.2.5). The estimator $\hat{\phi}$ is taken to be $\hat{f}_n$ according to lemma 3.1 and the finite set Θ_0 is the range of this estimator. To identify the global experiments $\mathbb{F}_i$ of the lemma, note that the measures in $\mathbb{F}_1(\phi)$ do not depend on ϕ (indeed $\mathbb{F}_1(\phi) = \mathbb{E}_{0,n-N_n}(f_0)$ is obtained by just restricting the parameter space in $\mathbb{E}_{0,n-N_n}$). Therefore $\mathbb{F}_1$ coincides with the set of product measures $P_f^{\otimes N_n} \otimes P_f^{\otimes (n-N_n)}$, $f \in \Sigma$, i. e. with $\mathbb{E}_{0,n}$. The experiment $\mathbb{F}_2$ coincides with $\mathbb{E}_{j,n}(\hat{f})$ as constructed; for $j = 3$ this again is a set

of product measures $P_f^{\otimes N_n} \otimes Q_{3,n-N_n,f}$. Take ϵ arbitrary; then for sufficiently large n we achieve (73) by theorems 2.1 and 2.8 (they were shown for sample size n; but since $n - N_n$ is of order n, the argument remains valid for the now relevant diminished sample size). We achieve (74) by lemma 3.1. We have shown $\Delta(\mathbb{E}_{0,n}, \mathbb{E}_{j,n}(\hat{f})) \leq 3\epsilon$ for sufficiently large n, which proves the theorem. □

11 Exact constants for L_2-risk

Proof of proposition 4.1. For this basic relation see Le Cam and Yang (1990), Strasser (1985), 49.6. These authors consider a setup of lower semicontinuous loss functions on a topological space of decisions. For our purpose it suffices to work with a measurable space of decisions $(G, \mathcal{G})$ and bounded loss functions $l_n(g, \vartheta)$ which are measurable in g. If $\mathbb{E}_i = (\Omega_i, \mathcal{A}_i, (P_{i,\vartheta}, \vartheta \in \Theta))$ is an experiment then (randomized) decision functions are Markov kernels $K : \Omega_i \times \mathcal{G} \mapsto [0, 1]$. The minimax risk is

$$\rho_i(l_n, \Theta) = \inf_K \sup_{\vartheta \in \Theta} \int l_n(g, \vartheta) K(\omega, dg) P_{i,\vartheta}(d\omega).$$

Proposition 4.1 is then immediate if both experiments $\mathbb{E}_i$, $i = 1, 2$ are Polish and dominated. Indeed, let $M : L(\mathbb{E}_1) \mapsto L(\mathbb{E}_2)$ be a transition attaining $\delta(\mathbb{E}_1, \mathbb{E}_2) + \epsilon$ and K be a decision function in $\mathbb{E}_2$. Since M is a Markov kernel (proposition 10.2), the composition $K \circ M$ is a decision function in $\mathbb{E}_1$, and we have for $\vartheta \in \Theta$

$$\int l_n(g, \vartheta) K(\omega_2, dg) P_{2,\vartheta}(d\omega_2) \geq \int l_n(g, \vartheta)(K \circ M)(\omega_1, dg) P_{1,\vartheta}(d\omega_1) - C\ (\delta(\mathbb{E}_1, \mathbb{E}_2) + \epsilon).$$

Taking a sup over $\vartheta \in \Theta$ and then an inf over K, we obtain, since $\epsilon > 0$ was arbitrary,

$$\rho_2(l_n, \Theta) \geq \rho_1(l_n, \Theta) - C\ \delta(\mathbb{E}_1, \mathbb{E}_2) \geq \rho_1(l_n, \Theta) - C\ \Delta(\mathbb{E}_1, \mathbb{E}_2).$$

In proposition 4.1 both experiments are Polish and dominated. □

Proof of proposition 4.4. For the Pinsker result many variants of proof have been given, see Golubev and Nussbaum (1990) (GN henceforth) and the literature cited therein. Our argument will therefore be extremely condensed.
(i): case $l(x) = x$. Set $q = [(n/\kappa)^r]$ for some $\kappa > 0$. Let

$$\breve{W}_2^\beta(\kappa) = \left\{ f \in \tilde{W}_2^\beta(\kappa),\ \int_0^1 f = 0,\ f^{(k)}(0) = f^{(k)}(1) = 0,\ k = 0, \ldots, \beta - 1. \right\}$$

Consider a probability measure ν on $L_2(0, 1)$ with finite support fulfilling $E_\nu \left\| D^\beta g \right\|^2 < \kappa$. Assume a prior distribution for f such that $f(x) = \sum_{k=1}^q n^{-1/2} q^{1/2} g_k(qx - k + 1)$ where g_k are i. i. d. ν. By the law of large numbers, this prior asymptotically concentrates on $\breve{W}_2^\beta(1)$ (lemma 5 in GN). By lemma 6 in GN, the minimax risk over $f \in \tilde{W}_2^\beta(P)$ with normed L_2-loss $n^{1-r} \left\| \hat{f} - f \right\|_2^2$ is then lowerbounded by κ^{-r} times the Bayes risk $\inf_{\hat{g}} E_\nu \left\| \hat{g} - g \right\|_2^2$ for prior ν in a model

$$dy(t) = g(t)dt + dW(t), t \in [0, 1] \tag{75}$$

(cp also Low (1993)). The set $\breve{W}_2^\beta(\kappa)$ has an ellipsoid representation, see section 5.1 of GN. Consider the Fourier coefficients $g_{(j)}$ of g wrt the pertaining orthonormal basis. Let $\Gamma \breve{W}_2^\beta(\kappa)$

the set of centered Gaussian distributions ν^* on $L_2(0,1)$ for which $g_{(j)}$ are independent and which fulfill $E_{\nu^*}\left\|D^\beta g\right\|^2 < \kappa$. Now ν may be selected to approximate such a ν^*, which yields a lower bound as a least favorable Bayes risk in the model (75)

$$\kappa^{-r} \sup_{\nu^* \in \Gamma \breve{W}_2^\beta(\kappa)} \inf_{\hat{g}} \int E\left\|\hat{g}-g\right\|_2^2 d\nu^*(g). \tag{76}$$

The eigenvalue asymptotics of the ellipsoid $\breve{W}_2^\beta(\kappa)$ is the same as for $\tilde{W}_2^\beta(\kappa)$; this implies that for $\kappa \to \infty$ the risk (76) tends to

$$\sup\left\{\int_{-\infty}^{\infty} h^2(x)\left(1+h^2(x)\right)^{-1} dx \ : \ \int_{-\infty}^{\infty} h^2(x)(2\pi x)^{2\beta} \le 1\right\}.$$

The value of the extremal problem is the Pinsker constant $\gamma(\beta)$ (cp. Golubev (1982)).
In this argument, since ν initially is a measure with finite support, the corresponding prior on f is such that almost surely

$$\sup_{x\in[0,1]} |f(x)| \le O((n^{-1}q)^{1/2}) = O(\kappa^{-r}n^{r-1})^{1/2}).$$

This proves that the lower bound remains valid with a restriction to $b_0(\tau_n)$.
(i): general $l \in \mathfrak{L}$. Combine the method in the lower bound proof of proposition 4.2 with the above argument.
(ii): **case** $l(x) = x$. Consider first the simple model of proposition 4.2, but assume now that the noise in (40) is $f_{0(j)}^{1/2}\xi_j$, where $f_0 \in \mathbb{R}_+^n$ is a vector such that $n^{-1}\left\|f_0\right\|^2 = 1$. Then for the estimator $\hat{f}_{(j)} = y_j/2$ we have for $f \in W_n$

$$\begin{aligned} E_{n,f}\, n^{-1}\sum_{j=1}^n (\hat{f}_{(j)} - f_{(j)})^2 &= n^{-1}\sum_{j=1}^n E_{n,f}(f_{0(j)}^{1/2}\xi_j/2 - f_{(j)}/2)^2 \\ &= \frac{1}{4}n^{-1}\sum_{j=1}^n f_{0(j)}^2 + \frac{1}{4}n^{-1}\sum_{j=1}^n f_{(j)}^2 \le \frac{1}{2}. \end{aligned}$$

i. e. risk performance of the optimal estimator $\hat{f}_{(j)}$ is the same as before, in the more general model with unequal $f_{0(j)}$. This phenomenon appears also in Pinker's ellipsoid model (43). In the more general model (44), consider the optimal estimator of proposition 4.3. It is known to be the minimax linear estimator over $\tilde{W}_2^\beta$ in (43), of form

$$\hat{f}^* = \sum_{j=0,\pm1,\pm2,\ldots} c_j \hat{f}_{(j)}\varphi_j$$

where $\hat{f}_{(j)} = \int \varphi_j dy$, for certain coefficients c_j , such that $c_j = c_{-j}$. The latter property holds since $\tilde{W}_2^\beta$ is symmetric wrt indices j and $-j$. For the risk of $\hat{f}^*$ in (44) we have (for each n only finitely many c_j are nonzero)

$$E_{n,f,f_0}\left\|\hat{f}^* - f\right\|_2^2 = \sum_j (1-c_j)^2 f_{(j)}^2 + n^{-1}\sum_j c_j^2 \int \varphi_j^2 f_0.$$

Observe that $\int(\varphi_j^2+\varphi_{-j}^2)f_0 = 2\int f_0 = 2$. Then $c_j = c_{-j}$ implies

$$E_{n,f,f_0}\left\|\hat{f}^* - f\right\|_2^2 = \sum_j (1-c_j)^2 f_{(j)}^2 + n^{-1}\sum_j c_j^2.$$

Thus we are back in the case of uniform variance function ($f_0 = \mathbf{1}$), where $\hat{f}^*$ attains the bound $\gamma(\beta)$ for $l(x) = x$.
(ii): general $l \in \mathfrak{L}$. Combine the method in the attainment proof of proposition 4.2 with the above argument.□

Proof of proposition 4.5. Consider the set

$$\mathcal{F}_n = \mathbf{1} + \tilde{W}_2^\beta \cap B_0(\tau_n).$$

In proposition 4.4 (i) $\tilde{W}_2^\beta \cap B_0(\tau_n)$ may be replaced by $\mathcal{F}_n$ since observations (and estimators) may be transformed by adding $\mathbf{1}dt$ to the observations $dy(t)$. Let $\beta \geq \alpha + 1/2$. We claim that τ_n may be chosen such that for any n

$$\mathcal{F}_n \subset \mathcal{W}_\epsilon^\beta,\ \mathcal{F}_n \subset \Sigma_n(\mathbf{1}). \tag{77}$$

Indeed, functions in $\mathcal{F}_n$ integrate to one. Furthermore they are eventually all $\geq \epsilon$ if $\tau_n \to 0$, so that $\mathcal{F}_n \subset \mathcal{F}_{\geq\epsilon}$ and the first inclusion is proved. By embedding theorems, $\tilde{W}_2^\beta(P)$ is contained in a Hölder class $\Lambda^\alpha(M)$ for $\beta \geq \alpha + 1/2$. Furthermore we have

$$\left\|\frac{f}{f_0} - 1\right\|_\infty = \|f - \mathbf{1}\|_\infty \leq \tau_n = o(\gamma_n)$$

for a choice $\tau_n = n^{-\beta/(2\beta+1)}\log n$ and $\beta > 1/2$, so that $\mathcal{F}_n \subset \Sigma_n(\mathbf{1})$. Since by remark 2.9 asymptotic equivalence holds over the set $\Sigma_n(\mathbf{1})$, the proof is complete. □

Proof of proposition 4.6. For $\beta > 1$, by embedding theorems $\tilde{W}_2^\beta(P) \subset \Lambda^\alpha(M)$ for some $\alpha > 1/2$, $M > 0$. Thus $\mathcal{W}_\epsilon^\beta \subset \Sigma$, and by theorem 3.2 we may pass to the compound Gaussian white noise experiment $\mathbb{E}_{2,n}(\hat{f})$, for a choice $N_n = n/\log n$ and a preliminary estimator $\hat{f}_n$. Consider the measures $R_{2,n,f}(\hat{f})$ and $Q_{2,n-N_n,f,f_0}$ as introduced in section 3. Take $\delta > 0$, and define $l^{(\delta)}(x) = l((1+\delta)x)$; then for sufficiently large n

$$l_n(g,f) \leq l((1+\delta)(n-N_n)^{1-r}\|g-f\|_2^2) = l^{(\delta)}((n-N_n)^{r-1}\|g-f\|_2^2) = l_{n-N_n}^{(\delta)}(g,f),$$

say. For any estimator $\hat{f}_n^\#$ in $\mathbb{E}_{2,n}(\hat{f})$ we have

$$\begin{aligned}\sup_{f\in\mathcal{W}_\epsilon^\beta}\int l_n(\hat{f}_n^\#, f)dR_{2,n,f}(\hat{f}) &\leq \sup_{f\in\mathcal{W}_\epsilon^\beta}\int\left(\int l_{n-N_n}^{(\delta)}(\hat{f}_n^\#, f)dQ_{2,n-N_n,f,\hat{f}_n}\right)dP_f^{\otimes N_n}\\ &\leq \sup_{f\in\mathcal{W}_\epsilon^\beta}\sup_{f_0\in\Sigma}\int l_{n-N_n}^{(\delta)}(\hat{f}_n^\#, f)dQ_{2,n-N_n,f,f_0}.\end{aligned}$$

Now take $\hat{f}_n^\#$ to be the estimator $\hat{f}_{n-N_n}^*$ of proposition 4.4 (ii), as a function of y in (37). Then uniformly over $f_0 \in \Sigma$

$$\sup_{f\in\mathcal{W}_\epsilon^\beta}\int l_{n-N_n}^{(\delta)}(\hat{f}_n^\#, f)dQ_{2,n-N_n,f,f_0} \to l^{(\delta)}(\gamma(\beta)).$$

according to proposition 4.4 (ii). Taking $\delta \to 0$ completes the proof. □

Proof of proposition 4.7. We first have to show that $\overline{\mathcal{W}}_\epsilon^\beta \subset \Sigma$. By embedding theorems we know that $f^{1/2}$ is in a class $\Lambda^\alpha(M')$ for $\beta \geq \alpha + 1/2$. Furthermore, the embedding inequality

$$\|g\|_\infty \leq C\left(\|g\|_2 + \left\|D^\beta g\right\|_2\right)$$

for $\beta > 1/2$ ensures that that $\left\|f^{1/2}\right\|_\infty$ is uniformly bounded. Hence

$$|f(x) - f(t)| = \left|f^{1/2}(x) - f^{1/2}(t)\right| \left|f^{1/2}(x) + f^{1/2}(t)\right| \leq 2CM\,|x-t|^\alpha,$$

hence $\overline{\mathcal{W}}_\epsilon^\beta \subset \Sigma$. By theorem 1.1 it now suffices to consider risk bounds in the white noise model (4). The attainability of the bound follows directly from proposition 4.3. Here it is to be noted that the factor $1/2$ of the noise appearing in (4) can be amalgamated into $n^{-1/2}$, i. e. into the normalizing n^{1-r} in (39). For the lower bound in (4) we have to take into account that $f^{1/2}$ is now restricted to the unit sphere in L_2. Let $b(t) = \{f;\ \|f\|_\infty \leq t\}$. We use proposition 4.4 (i) and further restrict $f^{1/2}$ to a set $\mathbf{1} + b(\tau_n)$ where $\tau_n \to 0$, $\tau_n n^{(1-r)/2} \to \infty$. Let $\Pi_1(f^{1/2})$ be the L_2-projection of $f^{1/2}$ to the affine tangent hyperplane of the unit sphere of L_2 at point $\mathbf{1}$. Then obviously $\Pi_1(f^{1/2}) = f^{1/2} + c_f \mathbf{1}$ for some number c_f, and

$$\begin{aligned} c_f &= \left\|f^{1/2} - \Pi_1(f^{1/2})\right\|_\infty = \left\|f^{1/2} - \Pi_1(f^{1/2})\right\|_2 \\ &= O\left(\left\|f^{1/2} - \mathbf{1}\right\|_2^2\right) = O(\tau_n^2) \end{aligned} \tag{78}$$

uniformly over $f^{1/2} \in \mathbf{1} + b(\tau_n)$. Since $n^{r-1} = n^{-2\beta/(2\beta+1)}$, and $\beta > 1/2$, τ_n may be chosen such that the r. h. s. of (78) is $o(n^{-1/2})$. We may then apply the reasoning in connection with (23) to show that in the white noise model where $f^{1/2} \in \mathbf{1} + b(\tau_n)$, the drift $f^{1/2}$ may be substituted by $\Pi_1(f^{1/2})$, with asympotic equivalence of the experiments. Then $\Pi_1(f^{1/2})$ varies in an affine subspace of L_2, and its derivative of order β for $\beta \geq 1$ coincides with that of $f^{1/2}$. Also (78) implies that by further restricting $f^{1/2}$, we can achieve that $h = \Pi_1(f^{1/2})$ varies fully within $\tilde{W}_2^\beta \cap (\mathbf{1} + b((1-\delta)\tau_n))$ for some $\delta > 0$. Subtracting $\mathbf{1}dt$ from the model yields a white noise experiment with parameter space $\tilde{W}_2^\beta \cap b_0((1-\delta)\tau_n)$, i. e. the case covered by proposition 4.4 (i).□

12 Addendum for proposition 1.2

Let Σ' denote an arbitrary set of probability measures on $[0,1]$. Define

$$S_n(\Sigma') = \left\{(P,Q) \in \Sigma' \times \Sigma' : H^2(P,Q) \leq n^{-1},\ P,Q \in \Sigma'\right\}.$$

and let $\frac{dP}{dQ}$ be the R-N- derivative of the Q-continuous part of P. Le Cam's second regularity condition for proposition 1.2 on the set of densities Σ is: if Σ' is the associated set of p. m. then

$$\sup_{(P,Q)\in S_n(\Sigma')} n(P+Q)\left(\left|\frac{dP}{dQ} - 1\right| \leq \epsilon\right) \to 0.$$

This is fulfilled in case $\Sigma' = (P_\vartheta, \vartheta \in K)$ where K is a compact subset of an open set $\Theta \subset \mathbb{R}^k$ and the family $(P_\vartheta, \vartheta \in \Theta)$ is differentiable in quadratic mean uniformly on compacts $K \subset \Theta$ (see proposition 1, chap. 17.3 in Le Cam (1986)).

Acknowledgement. The author wishes to thank David Donoho for encouraging discussions. Vladimir Koltchinskii, Mark Low and Enno Mammen suggested important improvements at various stages.

References

[1] Birgé, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrsch. Verw. Gebiete* **65** 181-237

[2] Brown, L. D. and Low, M. (1992). Asymptotic equivalence of nonparametric regression and white noise. To appear, *Ann. Statist.*

[3] Brown, L. D. and Zhang, C.-H. (1995). Nonparametric density estimation and regression are not asymptotically equivalent to the nonparametric white noise model when the smoothness index is $< \frac{1}{2}$. Ms.

[4] Donoho, D. L, Johnstone, I., Kerkyacharian, G., Picard, D. (1995). Wavelet shrinkage-Asymptopia ? *J. Roy. Statist. Soc. B* **57** No. 2 301-369

[5] Donoho, D. L., Liu, R. and MacGibbon, B. (1990). Minimax risk for hyperrectangles. *Ann. Statist.* **18** 1416-1437.

[6] Donoho, D. L. and Low, M. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944-970

[7] Dudley, R. (1989). *Real Analysis and Probability*. Wadsworth & Brooks/Cole, Pacific Grove, Cal.

[8] Efroimovich, S. Yu. and Pinsker, M. S. (1982). Estimating a square integrable probability density of a random variable (in Russian). *Problems Inform. Transmission* **18,** No. 3, 19-38

[9] Falk, M. and Reiss, R.-D. (1992). Poisson approximation of empirical processes. *Statist. Probab. Letters* **14** 39-48

[10] Golubev, G. K. (1982). On minimax filtering of functions in L_2 (in Russian). *Problems Inform. Transmission* **18,** No. 4, 67-75

[11] Golubev, G. K. (1984). On minimax estimation of regression (in Russian). *Problems Inform. Transmission* **20,** No. 1, 56-64

[12] Golubev, G. K. (1991). LAN in problems of nonparametric estimation of functions and lower bounds for quadratic risks. *Theory Probab. Appl.* **36,** No. 1, 152-157

[13] Golubev, G. K. and Nussbaum, M. (1990). A risk bound in Sobolev class regression. *Ann. Statist.* **18** 758-778

[14] Ibragimov, I. A. and Khasminski, R. Z. (1977). On the estimation of an infinite dimensional parameter in Gaussian white noise. *Soviet Math. Dokl.* **236**, No. 5, 1053-1055.

[15] Ingster, Yu. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives I-III. *Math. Methods Statist.* **2** 85-114, 171-189, 249-268.

[16] Koltchinskii, V. (1994). Komlos-Major-Tusnady approximation for the general empirical process and Haar expansions of classes of functions. *J. Theoretical Probability* **7** (1) 73-118.

[17] Korostelev, A. P. (1993). An asymptotically minimax regression estimate in the uniform norm up to an exact constant. *Theor. Probab. Appl.* **38** 4 737-743

[18] Le Cam, L. (1985). Sur l'approximation de familles de mesures par des familles gaussiennes. *Ann. Inst. Henri Poincaré* **21** (3) 225-287

[19] Le Cam , L. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer-Verlag, New York.

[20] Le Cam , L. and Yang, G. (1990). *Asymptotics in Statistics.* Springer-Verlag, New York.

[21] Low, M. (1992). Renormalization and white noise approximation for nonparametric functional estimation problems. *Ann. Statist.* **20** 545- 554

[22] Low, M. (1993). Renormalizing upper and lower bounds for integrated risk in the white noise model. *Ann. Statist.* **21** 577-589

[23] Mammen, E. (1986). The statistical information contained in additional observations. *Ann. Statist.* **14** 665-678

[24] Millar, P. W. (1979). Asymptotic minimax theorems for the sample distribution function. *Z. Wahrsch. verw. Gebiete*, **48**, 233-252

[25] Nikolskij, S. M. (1977). *Approximation of Functions of Several Variables and Imbedding Theorems.* Springer Verlag, Berlin, Heidelberg, New York, 1975.

[26] Nussbaum, M. (1985). Spline smoothing in regression models and asymptotic efficiency in L_2. *Ann. Statist.* **13** 984-997

[27] Parthasarathy, K. R. (1978). *Introduction to Probability and Measure.* Springer Verlag, New York

[28] Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission* (1980) 120-133

[29] Reiss, R.-D. (1993). *A Course on Point Processes.* Springer Verlag, New York

[30] Rio, E. (1994). Local invariance principles and their application to density estimation. *Probab. Theory Rel. Fields* **98** 21-45

[31] Shorack, G., Wellner, J. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.

[32] Strasser, H. (1985). *Mathematical Theory of Statistics.* Walter de Gruyter, Berlin.

[33] Tsybakov, A. B. (1994). Efficient nonparametric estimation in L_2 with general loss. Ms.

[34] Van de Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18** 907-924

[35] Woodrofe, M. (1967). On the maximum deviation of the sample density. *Ann. Math. Statist.* **2** 475-481

Weierstrass Institute
Mohrenstr. 39
D-10117 Berlin, Germany
e-mail nussbaum@wias-berlin.de